

The rise of café culture

A night out in a bar is all the more enjoyable if you can digest some science too. That's the lesson of a growing movement whose character may be local but whose reach is potentially global — and at a small cost.

Since last November, the small English market town of Settle in Yorkshire has added an unusual entry to its social calendar: a regular night of science debate. On a cold evening this February, for example, mathematician David Salinger journeyed from the nearby University of Leeds to discuss the mathematical meaning of infinity — hardly a typical night down the pub.

Salinger is not alone. In cities across Britain and France, academics regularly give up their evenings to join in with these Café Scientifique debates. And as a conference of debate organizers in Newcastle upon Tyne, UK, heard last weekend, similar events are now established in towns in eleven countries, from Buenos Aires to Warsaw to Houston, Texas (see page 333).

Given that the first Café Scientifique was held in 1997, the spread of the events has been remarkable and should be welcomed. Governments across Europe have stressed the need to improve dialogue between scientists and the public. The café organizers, using almost no public money and with little central organization, have created a network of successful events that does just that.

But a little central support could go a long way. The British cafés built a network thanks to a small grant — £175,000 (US\$315,000) — from the Wellcome Trust. The trust only provides money to kick-start such schemes, so another body will be needed to support the activities. This would surely be a worthwhile investment for any organization that cares about science communication.

The organizers also face a lingering distrust of their events from within the professional science-communication community. Sceptics point out that the events attract a mainly middle-class audience,

the type of people who already engage with science through museums and the media. A better event, they argue, would reach minority communities who feel excluded from science and, more importantly, from democratic decision-making about scientific issues.

This may be true, but it also misunderstands what the cafés can achieve. The events are run by local people who give their time free of charge. Event organizers know they are reaching a middle-class audience and, by and large, this does not bother them too much. These are not professional events — they are the science equivalent of the book club, run by enthusiasts, for enthusiasts.

Even if organizers were more committed to reaching minority communities, it is questionable whether the debate format is the best vehicle for doing so. Minority communities often ignore the mainstream science events, such as exhibitions and festivals, that already exist. Why should the cafés, developed and run by mainly white, middle-class professionals, be any different?

Inspiration should be drawn instead from programmes that have been created for particular communities. Take, for example, the Dana Centre in London, an offshoot of the Science Museum designed to host debates about science and culture.

Staff there have produced an event about climate change for the city's Bangladeshi population. The issue is relevant to the community, as rising sea levels could devastate their low-lying Asian homeland. Realizing that the community was unlikely to cross London to visit the museum, the centre took its event — an interactive theatre piece — to a local community centre. The result was a spirited debate and a determination by both sides to run more such shows. ■

Dragged into the fray

Willingly or not, Russia's science academy has become part of the political economy of climate.

We've been here before. Last December, Russian ministers and advisers were sending contradictory signals about whether they would ratify the Kyoto Protocol for climate change, in the end leaving the issue unresolved. Last week saw an apparently similar phenomenon, except that in the process the Russian Academy of Sciences apparently allowed itself to be hijacked by opponents of the protocol.

According to the news agency Reuters, a document handed to Russian President Vladimir Putin by the academy strongly opposed ratification, asserting that the protocol "lacks a scientific basis" and would put a brake on Russia's economic development. This matters, because Russia's ratification would bring the Kyoto Protocol into force.

The report, commissioned by Putin in January, was adopted by an academy commission on 14 May, despite disagreement among leading Russian climate scientists about its tone and recommendations.

Climate researchers are upset that the report's main author, Yuri Izrael, director of the academy's Institute of Global Climate and Ecology in Moscow and an influential scientific adviser to the Kremlin, is also a vice-chairman of the Intergovernmental Panel on Climate

Change (IPCC). As such, he was a signatory to the IPCC's 2001 assessment of climate change — the very document that flatly contradicts some of the Russian anti-Kyoto statements.

There are suggestions that the chairman of the IPCC, Rajendra Pachauri, might take some action against Izrael. This should be resisted, partly because the IPCC has better things to do, and partly because the Russian report may well be part of a political charade. It appeared in the same week as a summit between Russia and the European Union (EU) in which Putin sought approval and improved conditions for entry into the World Trade Organization. The EU is a strong supporter of Kyoto. Lo, Putin got the backing he wanted and, lo again, he immediately expressed his willingness for Russia to move towards ratification of Kyoto.

How and why the academy allowed itself to become a pawn may never be clear to outsiders. What is clear is that science in Russia, as elsewhere, has been hijacked by the politics and economics of energy investment and emission reductions. Anyone seeking to interpret messages about such science should apply the same filter of scepticism as they do to politics. Even so, the bottom line is that prospects for the Kyoto Protocol are brighter than they have been for a while. ■





Lost in space
'Creative' approach
to Beagle 2 finds no
favour with inquiry
p330



Risk factor
Tissue analysis hints
at broader spread of
brain disease
p331



Culture shock
Microbiologists get to
grips with growing
the world's bacteria
p332



Creature comfort
Britain sets out plan
to improve the lot of
lab animals
p334

Clinicians win fight to overturn patent for breast-cancer gene

Alison Abbott, Munich

A European patent that gave a company in Utah the exclusive right to perform diagnostic tests for a breast-cancer gene has been revoked.

In a landmark ruling on 18 May, the European Patent Office (EPO) granted an appeal against Myriad Genetics of Salt Lake City over its patent on the gene *BRCA1*. The office ruled that Myriad's claim was invalid because its original submission, made in the United States in 1994, contained a number of small errors in the gene's sequence.

The finding sheds light on the race in the early 1990s to patent genes of interest, while updating their accuracy, even as other researchers put versions of the same sequences in public repositories, such as GenBank.

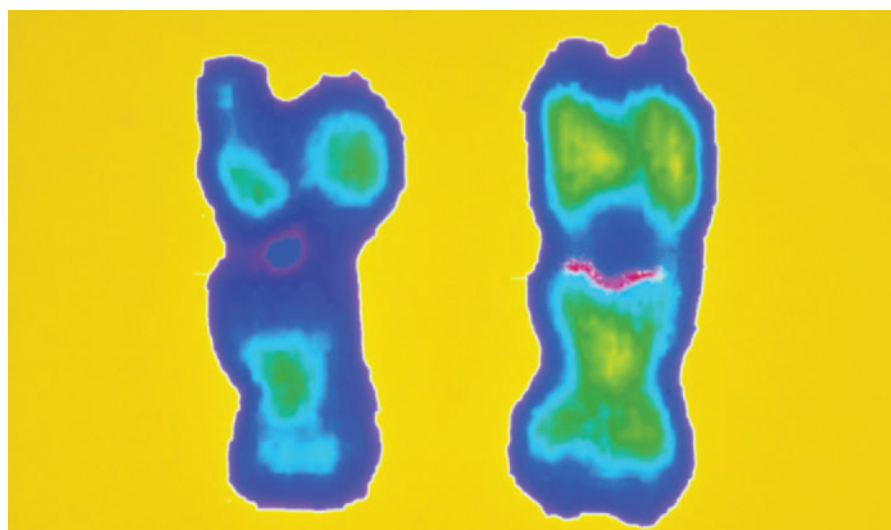
Mutations in *BRCA1* predispose women to some hereditary forms of breast cancer. Myriad's patent had incensed European clinicians because it allowed the company to require that all tests for the gene be done at its labs in Salt Lake City (see *Nature* **413**, 443; 2001). Many researchers refused to comply, continuing to use their own tests instead.

A group of opponents appealed against the patent, claiming that the identification of the gene owed too much to research in the public domain to justify Myriad's claim to have made an inventive step.

In 1990, several academic research groups formed an international consortium to home in on the gene, which had been located to a particular region of chromosome 17 (J. M. Hall *et al.* *Science* **250**, 1684–1689; 1990). The collaboration met regularly and pooled its data.

Once the region containing the gene had been narrowed down, several groups in the consortium, including some researchers from the University of Utah who went on to found Myriad Genetics, began competing against each other to identify the gene's sequence.

In its dash to secure rights on the gene, Myriad staked its first claim in a patent filed to the US Patent and Trademark Office in August 1994. Over the next twelve months, it filed an additional seven patents on the *BRCA1* gene with the US office, updating sequence information as it went along. It brought its claims



Women with only one copy of the *BRCA1* gene (red) on chromosome 17 may be at risk of breast cancer.

together in a single patent application to the EPO in August 1995. By this time, the gene's sequence was well known in the community.

In its EPO application, Myriad referred to its first US filing in August 1994 as its "priority date", but the examining officer spotted that the sequence, which encodes a protein made up of 1,864 amino acids, was shorter than it should have been. The officer insisted that Myriad change the priority date to that of its second US filing, in September 1994, failing to notice the less obvious differences between the sequence filed then and the correct one. In fact, ten amino acids dotted throughout the protein were incorrect.

Ruled out

These differences were only corrected in Myriad's fifth US application, filed in March 1995. By this time, others, including Myriad scientists, had published information on the sequence and so it was no longer novel, and thus no longer patentable, ruled the EPO's appeal board. This ruling will not directly affect Myriad's US patent.

"Finding such small differences in large genes is close to impossible," says Siobhan Yates, head of biotechnology at the EPO. Priority documents are always in a different format from EPO application documents, ruling

out electronic comparisons, she says.

Similar sequence differences have led to other EPO gene patents losing their priority dates, including patents claimed by Genentech in South San Francisco on tissue plasminogen activator, in which three nucleotides were incorrect, and the HIV virus, which had four errors. As a result, the scope of both patents was restricted.

"The EPO has sent a clear signal that it requires strict accuracy of information," says Jacques Warcoin, patent attorney for several of Myriad's opponents, including the Curie Institute in Paris, whose scientists spearheaded opposition to the patent.

But some opponents say that they wish the EPO had revoked the patent on matters of principle, rather than on technicalities. "Our concerns are about how the inventive process envisioned in patent law can be reasonably applied to disease genes," says Mike Stratton, a geneticist at the Sanger Institute near Cambridge, UK, who was part of the original consortium seeking the breast-cancer gene.

Myriad has two other patents covering *BRCA1* and its mutations, and opposition hearings for them have been scheduled for January 2005. It has until the end of the year to oppose the patent's revocation. Myriad did not respond to requests for comment. ■

Project structure blamed for Beagle 2 loss

Laura Nelson, London

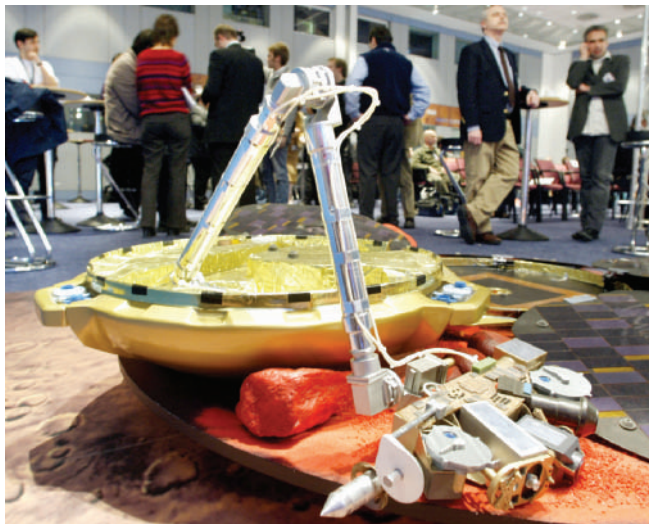
For the European Space Agency (ESA) and the British government, the Beagle 2 Mars lander was a bold experiment in the delegation of project management to the people building the probe.

But according to an official inquiry, this experiment was an expensive failure and should not be repeated. The £40-million (US\$72-million) probe was lost as it attempted to land on Mars on Christmas Day last year. Future projects, the inquiry says, should be far more rigorously controlled and documented.

The panel's summarized findings were released on 24 May and swiftly accepted by ESA. "The creative approach used with Beagle 2 was probably a step too far," admits David Southwood, the space agency's chief scientist. "Maybe I should have never let it go."

Beagle 2 was an adjunct to Mars Express, its mother craft, and ESA allowed its British design team, led by Colin Pillinger, a planetary scientist at the Open University, Milton Keynes, considerable autonomy in managing the project. The lander hitched a ride with Mars Express when it left Earth last June. When it reached Mars' orbit on 19 December, Beagle 2 was successfully ejected, but was never heard from again.

ESA contributed £20 million to the probe's costs and most of the rest came from the UK government. Additional funding came from EADS-Astrium — a Stevenage-based company that built some of the lander's components — the Open University and the National Space Centre in Leicester.



A mock-up of Beagle 2: the real thing remains missing after its trip to Mars.

The inquiry panel, which was chaired by ESA inspector-general, René Bonnefoy, and David Link, a former director of EADS-Astrium, argues forcefully that ESA should have treated Beagle 2 as an integral part of Mars Express.

The full report is not being made public, apparently to protect the confidentiality of submissions from ESA member governments and from private companies. But the panel released a list of 19 recommendations, which have been accepted by Southwood and by the British science minister, David Sainsbury.

About half of the recommendations concern the unprecedented level of autonomy that the project enjoyed from ESA's normal management structures. The panel roundly condemns this approach, and calls for future projects "which are critical to overall mission success or have a very high public profile" to be

managed directly by the agency, and subject to its normal planning and costing procedures.

The other findings highlight specific shortcomings in the testing and cost-accounting of the Beagle 2 mission. They say that future probes should have more robust telemetry and communications systems, should be rigorously tested for resistance to shock, and designed so that hatches and other detachable parts are less likely to collide with the main body if they fall off.

The panel also recommends far more rigorous testing of the parachute system, which some experts have identified as the most likely cause of Beagle 2's failure. But the inquiry reached

no firm conclusion on what became of the missing probe.

Accepting the findings at a press conference in London, Sainsbury said: "We must apply to the instrument the same kind of rigour and controls that are applied to the main spacecraft."

And Southwood accepted some personal responsibility for the craft's loss. "I was irresponsible not to have taken responsibility," he told *Nature*. He added that the probe's management structure had lacked accountability. "You've got to know who is giving orders to whom."

Pillinger said that the project team had done as well as it could. "We did our best, and that's all we could do," he said. "What I'm interested in is looking to the future," he added, urging ESA to make the earliest possible decision to return to Mars again. ■

Monsanto wins seven-year court battle for seed patent

David Spurgeon, Montreal

In a major boost for agricultural biotechnology, Canada's supreme court has ruled that a patent on a genetically modified seed extends to the cells and genes in the resulting plant.

The 21 May ruling is the last word in a seven-year battle between biotech firm Monsanto, based in St Louis, Missouri, and Saskatchewan farmer Percy Schmeiser.

Monsanto brought the action because its transgenic canola was found growing in Schmeiser's fields when he did not have a licence from the firm to cultivate the crop. Schmeiser, who maintains that the canola seeds blew on to his property, was ruled to have infringed Monsanto's patent by

growing the plants without a licence. The gene in question renders canola resistant to Monsanto's Roundup weed-killer.

The ruling, said to be the first of its type in any country, thrilled the biotech industry and angered consumer, environmental and farmers' groups. In a statement, Monsanto said that the ruling set "a world standard in intellectual property protection".

The case was seen as critical for the agricultural biotechnology industry, which holds thousands of patents likely to be affected by it and has many more pending. But critics say that the decision will give the industry too much control over farmers.

"This decision does not just condemn Percy Schmeiser, it also condemns the

broader community," says Andrea Peart, director of health and environment with the environmental group Sierra Club of Canada. "The responsibility of dealing with the environmental contamination by genetically engineered genes will now be shouldered by the public, not the polluter."

The ruling contrasts with the Supreme Court's 2000 decision on the Harvard Mouse, a transgenic rodent used in cancer research. It said that the mouse could not be patented in Canada, although courts in the United States and Europe have permitted patents. But the court found that the nature of commercial agriculture meant that plants grown from the seeds constituted a 'use' of the technology covered by the patent. ■

Tissue survey raises spectre of 'second wave' of vCJD

Jim Giles, London

Britain's epidemic of variant Creutzfeldt–Jakob disease (vCJD) could be worse than current death rates predict, a survey of a marker for the condition in body tissue suggests.

The disease, thought to be contracted by eating beef from cattle infected with mad cow disease, was first seen in Britain in 1995. Since then, 141 people have died, and government scientists have struggled to estimate the likely scale of the epidemic. vCJD can only be diagnosed after death, so previous estimates of its prevalence have been confined to projections from the known death statistics.

Now a study of around 13,000 tissue samples taken from hospital patients has found that three of them contained traces of a prion that is associated with vCJD (D. A. Hilton *et al.* *J. Path.* doi:10.1002/path.1580; 2004). The figure is surprisingly high: "Finding three cases in a sample this size is sobering stuff," says John Collinge, a prion-disease researcher at University College London.

The study, led by pathologist David Hilton of Derriford Hospital in Plymouth, looked at tissues from tonsils and appendices removed in hospitals across Britain. vCJD is believed to be caused by the build-up in the brain of an abnormal form of a normally benign but little-understood type of protein called a prion. Abnormal prions are known to accumulate in tonsil and appendix tissue in vCJD sufferers. Hilton's team first removed the normal form of the prion and then used an antibody that binds to either form to search for the abnormal version.

James Ironside, an author on the paper and director of the National CJD Surveillance Unit at the Western General Hospital in Edinburgh, says that current estimates of the disease's prevalence would predict that none — or perhaps one — of the samples would test positive. But he cautions that the numbers involved are too small to allow new estimates of overall vCJD prevalence. There is also an unexpected feature of the results: the distribution of the abnormal prion in two of the three positive samples was different from that previously observed in vCJD patients.

One explanation for the result is that more people are carrying the disease than the death rates suggest. All known individuals with vCJD have a particular genotype, called MM. People with other genotypes may incubate the disease for even longer, and so haven't yet shown up in the death rates. "There could be a second wave of cases," says Ironside. He would like to retest the two



The prevalence of vCJD in Britain remains a subject of contention.

positive samples that show odd distributions to see if they are from people with different genotypes, but isn't sure if there is enough tissue to do so.

Another possibility is that the survey has unearthed a group of people who are infected with abnormal prion proteins, but who will not go on to develop vCJD.

One prion-disease expert, who declined to be named, questioned whether Hilton's team should have gone public with their potentially alarming results on the basis of so few data. This researcher noted that false positives have emerged in similar surveys and says Hilton should have run western blot tests, which are more sensitive than the antibody staining method used by Hilton's group. Ironside says the method used by hospitals to preserve tissue samples — immersion in paraffin — precluded the use of western blot tests.

The pattern of vCJD infection in Britain should become clearer in late 2006, when the Health Protection Agency is due to complete a study of 100,000 tonsil tissue samples. These are frozen, and therefore suitable for western blot analysis.

Biodefence project accused of violating weapons treaty

Erika Check, Washington

Three prominent US arms-control experts have hit out at the Department of Homeland Security's plans to start a new programme in biodefence research.

In an article published online on 17 May, the specialists say that activities planned at a National Biodefence Analysis and Countermeasures Center, scheduled to be built at Fort Detrick in Maryland, could breach the Biological Weapons Convention. The convention, which the United States has signed, prohibits offensive bioweapons programmes.

According to a presentation made by a US Army official at a Department of Defense meeting on 9 February, the centre would carry out studies to anticipate how bioterrorists might develop and deploy new biological threats. But the authors of the article, published in *Politics and the Life Sciences* (www.politicsandthelifesciences.org), say that these studies may cross the line from defensive into offensive territory. Taken together, they write, many of the activities "may constitute development in the guise of threat assessment, and they will certainly be interpreted that way".

The authors are Milton Leitenberg, an arms-control analyst at the University of Maryland; James Leonard, who led US negotiations to the Biological Weapons Convention; and Richard Spertzel, a scientist who worked with the United Nations Special Commission investigating bioweapons in Iraq.

Leitenberg says he is also troubled by the fact that no independent group is reviewing all of the US government's planned biodefence research to decide whether it complies with the convention. He adds that a new body announced on 4 March, the National Science Advisory Board for Biosecurity, will not solve the problem because it is not permanent and will not review classified projects.

Most of the terrorist groups about which the United States is concerned, including al-Qaeda, have not developed effective bioterrorism programmes, Leitenberg adds. The United States risks accelerating their progress, he argues, by investigating new horizons in bioweapons. "We will become the Johnny Appleseed, providing sophisticated technical know-how related to bioweapons to the rest of the world," he says.

The Department of Homeland Security did not return calls seeking a response to the critique.

Microbiology gaining ground after lean years

Jonathan Knight, New Orleans

Of the thousands of microbes known to science, 99% have yet to be cultured in a dish. But unless they can be grown in the laboratory, there is little hope that their physiology or behaviour will ever be well understood.

Solving this culture problem was just one of the challenges outlined to microbiologists who attended a special, 24 May symposium on the future of their discipline at the annual meeting of the American Society for Microbiology in New Orleans.

With the rise of molecular biology over the past half-century — and that of biotechnology in its wake — researchers who study the life cycles and habitats of microbes have acquired something of a stuffy image. As research funding shifted to the fashionable areas of molecular and cell biology, microbiology faculty positions were lost, and in some cases whole departments were closed.

But now, the symposium was told, microbiology is regaining some of its lustre. Part of the revival is due to growing public concern about emerging diseases and biological weapons. Another contributing factor is the speed with which researchers are sequencing the full genomes of various microbes.

Almost 200 bacteria and fungi have been completely sequenced to date. This makes microbes particularly useful for answering big-picture questions about biodiversity and evolution, says Moselio Schaechter, a mushroom biologist and visiting scholar at the University of California, San Diego. "We are entering a golden age of microbiology," enthuses Schaechter, who moderated a meeting in Charleston, South Carolina, last September to identify growth areas in the field.



Popular culture: the study of bacteria and viruses is experiencing a resurgence.

These sequences offer new opportunities for microbiologists. They could aid the study of gene sharing, an important factor in evolution: bacteria are the most promiscuous gene exchangers known — up to 60% of the genes in the gut bacterium *Escherichia coli* can be different in different strains. No one yet knows how, or why, this is so.

Rapid sequencing enables entire communities of microbes to be identified from a single sample of water, as shown in projects such as the study of acidic run-off from mines run by

the Joint Genome Institute in Walnut Creek, California. The approach ensures that all microbial species in a sample can be found, not just those that grow readily in culture.

"This will be a key part of how we look at complex populations in the future," predicts Stanley Malloy, director of the Center for Applied and Experimental Genomics at San Diego State University and president-elect of the microbiology society. By monitoring genomes of entire populations, microbiologists hope to answer perplexing questions, such as why bacterium-infecting viruses that live in the ocean carry genetic instructions for synthesizing cholera toxin. "Why are they there?" asks Malloy. "Nobody knows."

One sure sign that times are changing is that microbiologists are finding jobs in university departments outside biology, says Roberto Kolter, who studies sheets of bacteria, known as biofilms, at Harvard Medical School in Boston. Earth sciences departments hire them to study the microbial processes that shape rock, and astronomy departments want help from experts on microbes living in extreme environments to investigate the possibility of life on other planets.

Harvard, which lacks a microbiology department, established a Microbial Sciences Initiative earlier this year that will create faculty positions for microbiologists elsewhere in the university. Although their primary labs will still be distributed around campus, the faculty members will meet regularly and share some laboratory space.

Regardless of the form microbiology takes, participants at the meeting expect that it will make a big contribution to our understanding of life on Earth. "Microbes rule this world," Schaechter says. ■

Japan announces follow-up to human genome project

David Cyranoski, Tokyo

Japan is launching the Genome Network, a five-year, ¥15-billion (US\$130-million) initiative that will attempt to build on the human genome project and systematically study the function of all human genes.

The initiative will collate experimental data on all 30,000 human genes, says Yoshihide Hayashizaki, a researcher at the Genomic Sciences Center (GSC) in Yokohama and member of the project's steering committee. The GSC will play a central role, but Japan's science ministry will divide the work among researchers across the country. Organizers plan to start recruiting teams next week.

Details of the initiative remain to be decided. Like the US Encyclopedia of DNA

Elements (ENCODE) project, it aims at a systematic, comprehensive analysis of the human genome. Japan's project will focus on how the information in the genome is expressed, says Yoshiyuki Sakaki, who is head of the GSC and also on the steering committee. It will look at the interaction between genes and the proteins known as transcription factors that initiate expression of them, he explains.

The project hopes to assemble a library of complementary DNA corresponding to every human gene. The collections of cDNA held by Hayashizaki's centre and by Sumio Sugano of University of Tokyo's Institute of Medical Science are already among the largest in the world, notes Sakaki. "We need to take advantage of this," he says.

Takehiko Sasazuki, head of research at the International Medical Center of Japan in Tokyo and chairman of the steering committee, says the scale of the project will help ensure its success. "By putting together all these data, we will be able to find unexpected phenomena."

Sakaki and others say that the initiative will begin at a national level but may collaborate with projects in other countries later. "First we have to run our own system and then we can figure out how to cooperate," says Sakaki. "We need a year."

He adds that details of the project will be agreed at a final steering committee meeting on 31 May, and that work will begin in September, with funding already assured from the science ministry. ■



Sweet talk: Tom Shakespeare, a prime mover at *Café Scientifique*, engages an audience in Bologna, Italy.

Pop science pulls in public as café culture goes global

Jim Giles, Newcastle upon Tyne

The Vitenskapskafeen in Oslo is so popular that people happily sit on the floor once the seats are taken. The same kind of event, when held in a forest in Poland, attracts everyone from villagers to the local priest. In Denver, Colorado, such meetings draw enough singles to create match-making occasions.

The focus of each of these popular gatherings, called *Cafés Scientifique*, is informal scientific debate. The idea was started by French physicists in 1997, but when *Café Scientifique* organizers met to share ideas in Newcastle upon Tyne, UK, on 21 and 22 May they welcomed colleagues from 11 different countries, from Argentina to Japan.

The meeting revealed a movement that has expanded rapidly, but remains close to its grass roots. Suggestions that international cafés could develop a more unified structure and organize larger sources of funding attracted limited interest. Instead members were keen to learn small tricks of the trade from each other.

The events are all based in informal settings, such as bars and cafés. Expert speakers, such as scientists from local universities or businesses, start the meeting with a short talk before joining an audience-led debate. Organizers deliberately keep audiences small, aiming for around 50 participants.

In Britain and France in particular, the proliferation of *Cafés Scientifique* has been dramatic — the two countries now host around 70 regular events a year between them.

Many of the organizers are eager to see what else they can do with the format; they could, for example, use video conferencing for debates between disparate groups of people. Michael Puncheon, who runs café events as part of his work at the British Council, describes one successful event that

linked scientists and lay people in London, Istanbul and Ramallah, on the West Bank near Jerusalem. They were given the chance to hear Mike Parker, an expert in the ethics of genetics at the University of Oxford, discuss issues such as genetic testing. “They were young audiences and they really got stuck in,” says Puncheon.

The idea of cafés for young people also attracted interest at the meeting. Since the first such event was run in 1999 in Lyon, France, junior debates have spread to schools in seven French cities. UK organizers say that in the autumn they will apply for funding for a similar programme to the Wellcome Trust, a research charity that encourages public engagement with science. “It’s a chance for pupils to meet scientists and businessmen — the people who shape the world,” says organizer Pablo Jensen, a physicist at Claude Bernard University in Lyon.

Tom Shakespeare, a sociologist at the University of Newcastle upon Tyne who helped organize the meeting, acknowledges that the cafés appeal mainly to the middle class. But Shakespeare and others point out that the events still reach a wide audience, particularly as they can be held in small towns that do not have science museums or universities to host larger events. The cafés also serve an educational function: in France, events have been run in teacher training colleges to make new teachers more confident about discussing science.

The Denver café, organized by University of Colorado immunologist John Cohen, performs another social function: “I know that people have got together after meetings,” he says. “So when people come in alone, we carefully direct them to potential partners.” ■

See Editorial, page 327

■ www.cafescientifique.org

Britain opens first repository to speed work on stem cells

Laura Nelson, London

Embryonic stem-cell research in Britain is expected to gain impetus from the opening on 19 May of a purpose-built stem-cell bank, just north of London. But despite generous funding and a regulatory framework carefully crafted to encourage the studies, researchers still face obstacles.

Britain established a framework to permit work on stem cells in 2001, but research has moved more slowly than some in the field anticipated. However, biologists are confident that work will go ahead, now support systems are in place.

Britain’s approach contrasts with that of the United States, where work on only certain embryonic stem-cell lines gets federal funding. But the difference has yet to set off a flurry of activity in Britain — or to spark a ‘brain drain’ of US cell biologists, scientists say.

“Stem-cell research is definitely progressing,” says Stephen Minger, a researcher in the field at King’s College London, “but it’s incredibly difficult work. I think the bank will accelerate things.”

Minger is one of the first to donate an embryonic stem-cell line to the new bank, which will house embryonic, adult and fetal stem-cell lines, and is located at the National Institute for Biological Standards and Control, Potters Bar.

The bank will make cell lines available to researchers worldwide, and provide British scientists with technical assistance in making use of the lines. Research agencies have allocated £40 million (US\$72 million) over three years to support its work.

Anne McLaren, who works on mouse stem cells at the Gurdon Institute in Cambridge, says that researchers are still struggling with the technical and regulatory challenges of stem-cell work.

Before deriving cell lines, researchers must obtain a licence from the Human Fertilisation and Embryology Authority, which takes about a year. Ten research groups have obtained these so far. After that, cell lines take several months to derive. “It does require a certain determination to get through the process,” says Robert Terry, a policy adviser at the Wellcome Trust, Europe’s largest research charity.

Austin Smith at the Institute for Stem Cell Research in Edinburgh, who obtained the first licence to derive embryonic stem cells, thinks activity in the field will soon pick up. “Things are starting in a serious way,” he says. ■

Homeopathy group loses lawsuit against TV science show

Munich An Italian court has declared homeopathy to be “not a serious therapy” and “substantially a medicine of emotions”, at the end of a four-year legal battle between homeopaths and an Italian television host.

In July 2000, guests on a popular-science television programme called *Superquarks*, hosted by Piero Angela, described how homeopathy could, if used in place of conventional therapies, constitute a risk for patients with serious or progressive illnesses. They also indicated that clinical benefits are due to a placebo effect. The Italian Association of Medical Homeopathy sued Angela for failing to represent its viewpoint in the broadcast (see *Nature* **412**, 261; 2001).

In a 59-page judgement on 26 April, the court in Catania decided that the opinions of the guests were “justified” and that the message of the programme “could not be considered offensive or defamatory”.

Money doubled to reform lab animal use in Britain

London The British government has backed its commitment for improvements in

animal testing with a rise in funding.

After weeks of speculation, UK science minister David Sainsbury announced on 21 May that money for the replacement, refinement and reduction of animal use — sometimes referred to as the ‘three Rs’ — will double from £330,000 (US\$600,000) last year to £660,000 this year.

The funds will be put towards a virtual national centre, coordinated by the Medical Research Council’s Centre for Best Practice for Animals in Research. The money will go towards developing alternatives to animal testing, such as computer modelling programs, to improve laboratory conditions, and to design research strategies that use fewer animals.



Pressing for change: Britain is trying to find ways of using fewer rats in lab experiments.

California dreaming of keeping nuclear labs

San Diego Faculty members at the University of California want the university to continue managing two nuclear-weapons labs, according to a poll by the university. At the same time, they strongly oppose a recent push for one of those labs to manufacture weapons components.

Los Alamos National Laboratory in New Mexico and Lawrence Livermore National Laboratory in California, which oversee the US nuclear-weapons stockpile, have been under the university’s management for decades. But, following a series of security lapses at the facilities last year, the government decided to put the operational contracts up for bidding.

The University of California has not yet decided whether it should enter the competition. Several other universities, including the University of Texas, as well as industrial firms, are eyeing the contract.

About a quarter of the 12,000 polled faculty members responded, of whom 67% were in favour of continued management. “The faculty has given a clear message about our desire to retain the labs,” says George Blumenthal, vice-chairman of

the academic senate for the ten University of California campuses.

AIDS-vaccine firm needed by stock-exchange rules

Washington The biotechnology company that carried out the first large-scale clinical trials of an AIDS vaccine has been warned that it could be delisted from the NASDAQ stock exchange.

VaxGen, which is based in California, saw its AIDSVAX vaccine fail clinical trials in both February and November of 2003. Since then, VaxGen has shifted its focus and won contracts from the US National Institutes of Health (NIH) to develop a vaccine against anthrax.

As *Nature* went to press, the company was late filing its quarterly financial reports. VaxGen says it delayed filing while studying whether it can count its NIH funding as revenue before reaching certain contract milestones. This would allow VaxGen to decrease its losses to shareholders from the past two years, which have been reported at \$74.4 million.

VaxGen has requested a hearing with NASDAQ, which will stay the delisting past its 27 May deadline for the financial report. The company says it is working towards filing as soon as possible.

Taiwanese satellite hunts for elves and sprites

Tokyo Researchers will soon be treated to new pictures of red and blue lights dancing above the clouds. On 20 May, Taiwan's National Space Program Office launched a satellite to take pictures of rare events called jets, elves and sprites — colourful discharges of light between thunderclouds and the ionosphere, at an altitude of 15 to 90 kilometres (see picture).

Previous studies of these transient events have used ground-based observations (see *Nature* **423**, 974–976; 2003). Satellite images of the flashes will help researchers to understand the global electric circuit, giving them better pictures in poor weather conditions and without the visual interference caused by the upper



atmosphere. "With the satellites we can determine the global distribution of the events and find new characteristics of them," says Nan-Hung Ting, who heads the office's science research programme.

Europe opens the door to transgenic corn

London An unofficial moratorium that blocked the approval of genetically modified foods in Europe came to an end last week.

On 19 May, the European Commission issued a licence for the sale of fresh or canned transgenic *Bt11* maize (corn). The maize, grown primarily in the United States and marketed by Swiss-based firm Syngenta, contains a bacterial gene that protects the crop against pests. It is the first transgenic product to be approved

for sale in Europe since 1998.

But trade experts say that the small step is unlikely to resolve transatlantic disputes over transgenic products. The World Trade Organization is considering a complaint brought by the United States in May last year (see *Nature* **423**, 369; 2003), which claims that the moratorium is illegal. Trade analysts say that US officials may also lodge a complaint about Europe's labelling policy, which requires that all food containing at least 0.9% transgenic content is labelled as genetically modified (see *Nature* **424**, 116; 2003).

Striking back

Stroke has disabled millions of people, stealing their ability to walk or communicate. Can future victims be helped by treatments that stimulate the growth of new brain cells? Alison Abbott investigates.

It is cruel, capricious and a killer. In the United States alone, about 700,000 people suffer from stroke every year, and up to 20% of them die within three months. Half are left disabled — many robbed of their power of speech, or their ability to walk or to manipulate objects.

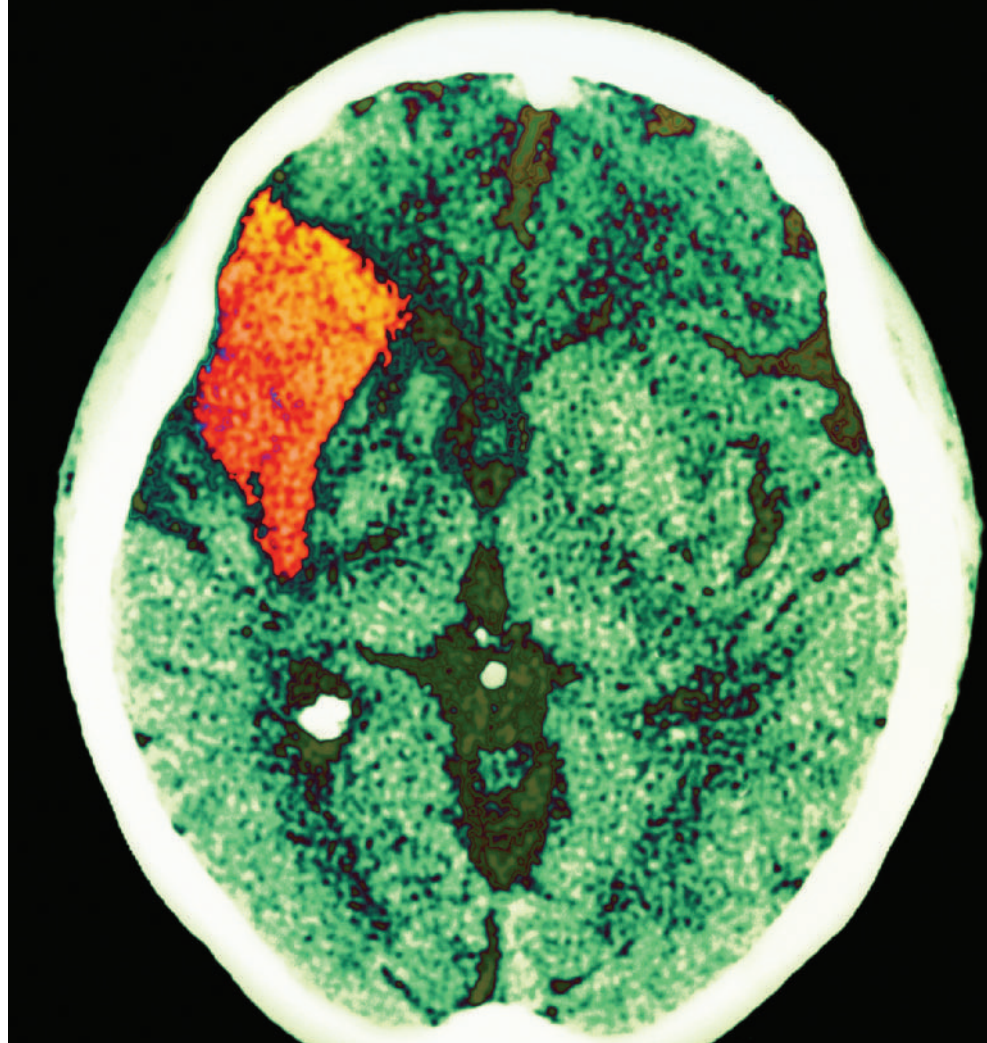
Nevertheless, almost all stroke survivors regain at least some of their lost function over time. This is mostly achieved by their brains recruiting other neural circuits to take over the function of those that were damaged during the stroke, when a blood clot or a ruptured vessel starved part of the brain of oxygen. But in recent years, evidence has emerged that injured brains can also grow new neurons — contrary to decades-old biological dogma. And that has set neurologists thinking: suppose this natural mechanism could be supercharged, so that a stroke-damaged brain repaired itself.

The idea arose after neural stem cells¹ were found in the adult brain. These can divide and differentiate into the two main classes of brain cell: neurons, which relay neural signals, and astrocytes, which provide a supportive environment for neurons to do their job.

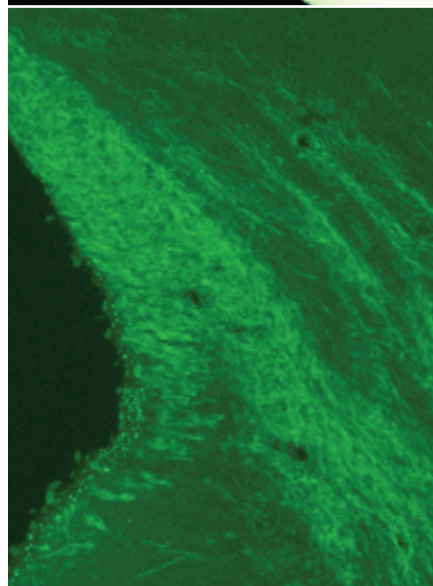
Self-help system

Normally, the turnover of these stem cells is vanishingly low, and most of the differentiated cells produced are astrocytes. But in 2001, evidence emerged that traumatic brain injuries, such as those caused by a stroke, can stimulate stem cells to generate new neurons, at least in rodents². These neurons have since been shown to migrate to the site of damage³⁻⁵. Left to itself, this mechanism is not up to the task of repairing gross damage caused by a stroke. But given the right tweaks and prompts, could it be made more efficient?

Stem cells capable of differentiating into neurons seem to exist mainly at two sites in



Repair mechanism: this brain scan shows the extent of cell death (above, orange) caused by a stroke. Researchers have now discovered that, contrary to traditional belief, neural stem cells in the brain can generate replacement neurons (left, green) that will migrate to some injured sites.



Jack Parent of the University of Michigan Medical Center in Ann Arbor independently studied a rat model of stroke in which a brain area called the striatum, which lies close to the SVZ, was damaged by briefly blocking blood flow in the middle cerebral artery. The resulting damage stimulated the growth of new neurons in the SVZ, which travelled up to the damaged areas^{3,4}.

Limited response

But the replacement of damaged cells was not extensive. Lindvall, for example, found that after a few weeks only 0.2% of the cells that had been destroyed were replaced, and that as many as 80% of the new neurons that had been produced had died³. "But it showed that the self-repair response is there," he says.

At about the same time, Masato Nakafuku, then at the University of Tokyo, reported on his results using a different rat model of stroke. Nakafuku's team damaged the hippocampus by briefly interrupting the entire blood flow to the brain, and found that new neurons were produced in a nearby area, which migrated into the damaged region⁵.

the adult brain. One is the subventricular zone (SVZ). In rodents, this feeds new neurons into the olfactory bulb to continuously retune the sense of smell. The other site is near the hippocampus, a brain region involved in learning and memory. There it continuously feeds a low level of new neurons into the hippocampus, particularly during active learning^{6,7}. Stem cells are also dotted around other parts of the brain, but they tend to differentiate into astrocytes.

In 2002, teams led by Olle Lindvall of Lund University Hospital in Sweden and



Physiotherapy (left) can aid recovery after treatment for a stroke (right), but do new neurons help too?

Most encouragingly, the group also showed that infusions into the brain of two growth factors, called EGF and FGF2, which cause stem cells to proliferate, greatly increased the number of new neurons produced — by almost ten fold in the case of EGF.

These new neurons formed connections with existing neuronal circuits. And the animals treated with growth factor seemed to recover some of the learning skills that they had lost immediately after their strokes. Nakafuku, who is now at the Cincinnati Children's Hospital Medical Center in Ohio, can't say for sure that this recovery was a result of the connections made by the new neurons. But a recent study, which has shown that irradiating rats to destroy stem cells before inducing stroke reduces the recovery of learning⁶, suggests that it may well be.

Encouraged by these results, many groups are now trying to enhance these effects. "Both the timing and duration of growth-factor infusion after stroke is critical," Nakafuku suggests. Parent adds: "We want to see if new, stroke-induced neurons integrate and function properly." That is an important issue, as the neurons that grow following epileptic seizures can integrate improperly into the brain's neural circuits,

and actually help to ensure that symptoms continue⁷.

Researchers want to know why so few of the neurons generated after a stroke survive long-term — which will require a better understanding about the environment of the stroke-damaged brain. The inflammation associated with stroke could be to blame, says Lindvall. Certainly, inflammation impairs the normal turnover of new neurons in the hippocampus^{8,9}.

It will also be important to induce the production of neurons in regions of the brain other than the striatum and hippocampus. So far, researchers have had little success in replacing neurons in the brain's cortex, which is frequently damaged by stroke. "But this may just be because we don't yet know how to do it," says Lindvall. Several research groups are now trying to develop animal models that will allow the cortex to be studied more precisely, he says.

David Greenberg of the Buck Institute in Novato, California, believes that important biological clues could come from studying animal models of neurodegenerative diseases such as Alzheimer's and Huntington's, in which it is now emerging that some of the dying neurons are replaced^{10,11}. "Stroke is an acute insult and all sorts of things spill

"Antidepressants, cholesterol-lowering drugs and even Viagra have all been shown to induce the growth of new neurons."

out as cells die en masse," Greenberg says. By studying the more sedate process of neuron damage and replacement that occurs in neurodegenerative diseases, he argues, it should be easier to understand the biological factors involved.

Human touch

Although much of the research into brain repair will continue to be done in rodent models, scientists would also like to observe the production of neurons in people. Several labs are trying to develop cell markers that could allow the differentiation of stem cells and the migration of new neurons to be tracked using positron emission tomography brain imaging. Such markers are likely to be structurally related to the building blocks of DNA, and so will be incorporated as dividing cells copy their genetic material.

In parallel to the work on basic biology, therapeutic leads are accumulating. Many different growth factors, in addition to EGF and FGF2, are being tested for their ability to promote the production of nerve cells. And because growth factors are proteins, which are difficult to administer as drugs and would have to be infused directly into the brain, researchers are also searching for smaller molecules with similar effects.

"There are plenty of candidates," says Alex Mercer, head of target discovery at the company NeuroNova in Stockholm, Sweden. Some therapeutics currently in use may provide a starting point for drug development. It is unclear whether these drugs will be as potent as growth factors, but antidepressants¹², the cholesterol-lowering statins¹³ and even Viagra¹⁴ have all in the past couple of years been shown to induce the growth of new neurons.

Given the vast amount of work to be done on basic biology and drug development, researchers warn that it will be many years before hopes of a revolution in stroke therapy might be realized. "It is not that we are pessimistic," says Lindvall. "It's just that we need an awful lot of time to do the basic research that is necessary before we can think of approaching patients."

Alison Abbott is Nature's senior European correspondent.

1. Gage, F. H. *Science* **287**, 1433–1438 (2000).
2. Jin, K. *et al. Proc. Natl Acad. Sci. USA* **98**, 4710–4715 (2001).
3. Arvidsson, A., Collin, T., Kirik, D., Kokaia, Z. & Lindvall, O. *Nature Med.* **8**, 963–970 (2002).
4. Parent, J. M., Vexler, Z. S., Gong, C., Derugin, N. & Ferriero, D. M. *Ann. Neurol.* **52**, 802–813 (2002).
5. Nakatomi, H. *et al. Cell* **110**, 429–441 (2002).
6. Raber, J. *et al. Ann. Neurol.* **55**, 381–389 (2004).
7. Parent, J. M. *Neuroscientist* **9**, 261–272 (2003).
8. Ekdahl, C. T., Claassen, J.-H., Bonde, S., Kokaia, Z. & Lindvall, O. *Proc. Natl Acad. Sci. USA* **100**, 13632–13637 (2003).
9. Monje, M. L., Toda, H. & Palmer, T. D. *Science* **302**, 1760–1765 (2003).
10. Curtis, M. A. *et al. Proc. Natl Acad. Sci. USA* **100**, 9023–9027 (2003).
11. Jin, K. *et al. Proc. Natl Acad. Sci. USA* **101**, 343–347 (2004).
12. Santarelli, L. *et al. Science* **301**, 805–809 (2003).
13. Chen, J. *et al. Ann. Neurol.* **53**, 743–751 (2003).
14. Zhang, R. *et al. Stroke* **33**, 2675–2680 (2002).

Catch a falling star

When NASA needed someone to grab a spacecraft on its re-entry to Earth, it turned to Hollywood for help. Nicola Jones meets the stunt pilots preparing to snag the \$260-million experiment before it hits the desert floor.

NASA/JPL/CALTECH

The ground is shaking and the rocky outcrops are crumbling as the volcano prepares to blow. The two trapped volcanologists, one injured by falling rubble, need to be rescued. Fortunately, Cliff Fleming arrives in his helicopter and plucks Pierce Brosnan and his co-star to safety in the filming of *Dante's Peak*. Years earlier, it was Fleming who dropped Bruce Willis's stunt man from his helicopter onto the wing of a jet plane in *Die Hard 2*.

But this September, Fleming will face a stranger challenge than anything he's done for the movie business, when he has a real star (or at least part of one) hanging from his helicopter. It is his job, along with a crack team of scientists, engineers and flight experts, to pluck a NASA spacecraft from the sky as it glides down to Earth, carrying with it a few precious grains collected from the Sun.

"When I got the call I thought, are you kidding? How'd you get my number?" recalls Fleming of the day NASA got in touch in the late 1990s. As owner of South Coast Helicopters in Costa Mesa, California, he was used to calls from film and construction crews — not from space agencies. "I thought the craft would be a lot bigger and burning up. I said: 'I don't think so,'" he laughs. It turns out they were serious.

So serious that we are now standing with eyes and binoculars glued to the sky on a cold mid-April morning, waiting for the pilots to do their stuff. The salt flats of the military training range outside Salt Lake City, Utah, stretch away in all directions, offering no protection from the freezing wind whistling around our ears. Along with Fleming and a dozen scientists and engineers, I watch as a tiny black speck appears — a helicopter 4 kilometres up.

As it flies over us, a paragliding capsule drops through the clouds, and the chopper catches it neatly on a hook and line. Fleming's fellow stunt man, Dan Rudert, has completed a successful test run. "He makes it look easy," smiles Fleming, as Rudert gently places the 1.5-metre, 190-kg mock capsule on the desert floor.

On 8 September, Rudert and Fleming will be up in the air to catch the real thing: the Genesis capsule. Early that morning, the capsule will separate from the rest of the spacecraft, just before entering Earth's atmosphere, and drop into the skies above Utah at about 9 a.m., with only the two pilots between it and a very bumpy landing.



It seems an oddly sensationalist way for NASA to go about its business. But watching the pilots in action it is clear this is no reckless stunt. For Don Burnett, a geochemist from the California Institute of Technology in Pasadena and the principal investigator behind the project, this is the surest way for the Genesis mission to make it back home safely. He would hardly settle for anything less — after all, the capsule is bringing back the first samples from space since the Apollo missions to the Moon 32 years ago. Its precious cargo — 400 micrograms of particles plucked from the solar wind — will give scientists much-needed clues to the early make-up and history of our Solar System.

Glowing report

The Sun dwarfs every other object in the Solar System, yet we know little about its precise composition. We know from spectroscopic measurements of its surface that it is mostly made of hydrogen and helium. And spacecraft observing the solar wind — a stream of ionized particles flowing from the outer layer of the Sun at hundreds of kilometres per second — show that it too is made of these light elements. But Genesis scientists are keen to know more about a sprinkling of other elements, such as oxygen and nitrogen,

that make up just 1% of this stellar outburst.

One mystery that they hope to solve is why different parts of the Solar System have such wildly different proportions of oxygen isotopes. These signatures are so distinctive that when a meteorite lands on Earth, scientists can use them to tell where it came from. As most space scientists believe that the Sun, planets, moons and asteroids were all created eons ago from the same massive cloud of gas and dust — the solar nebula — this lack of uniformity is intriguing.

The solar wind is a crucial part of this puzzle. With very few nuclear or chemical reactions going on in the outer layer of our star, the material here is essentially the same as when the solar nebula first collapsed 4.6 billion years ago. Measure the quantities of isotopes from this part of the Sun as they stream out in the solar wind, and you know what the Solar System looked like before a star or planets emerged. Today, we know these numbers with up to ten times less accuracy than we do for rocks from the Moon, Earth or Mars. Not good enough for scientists such as Burnett, who wants to compare those numbers with each other to work out which theories about the birth of our Solar System are right.

Although the arrival of the samples is much anticipated, their safe return is not yet



Caught in the act: stunt pilot Dan Rudert (above) completes a successful test run for catching the Genesis probe. In September, the events will happen for real, and Rudert and fellow pilot Cliff Fleming (top left) will attempt to prevent the first samples returning from space for 32 years from being destroyed on impact with the ground.

satellites. And military helicopters have caught reconnaissance drones weighing up to 1,800 kilograms over Vietnam.

The military pilots picked up the parachutes by flying straight at them, just missing the target. "It was nearly a mid-air collision," says Bob Corwin, Genesis's recovery operations team chief and an engineer from Lockheed Martin in Littleton, Colorado. "Scary from the cockpit."

In the 1990s, someone had the bright idea of using a parafoil instead of a chute. Like a paraglider's wing, these are strips of silk that inflate and drift with the air currents. That makes it much easier for the pilots, who can sneak up behind their prey instead of having to plough straight into a vertically dropping object. "It's like trying to go in formation with a slow-swimming jellyfish," says Roy Haggard, chief executive of the California-based aerospace company Vertigo, and a paragliding enthusiast who developed the technique for the US Navy in the 1990s.

When the Genesis capsule designers asked Haggard about the best method of mid-air recovery for their mission, he tracked down Fleming, one of a handful of stunt pilots worldwide who has experience in 'long-lining' — hanging a load off the back of a chopper. "It's a specialized skill," says

Fleming, sitting back in his chair in the hangar of the Dugway Proving Ground, part of the Utah Test and Training Range where the Genesis crew has been flying test runs for the past week. "Most pilots sit up and look out the front. But we have to fly with our heads stuck out the window, looking down." The helicopter in the hangar has no doors, and no instruments to tell the pilot how far away the target is or whether, once caught, it's beginning to swing. It's all done by feel.

In 2003, the team brought in Rudert, another long-liner and Hollywood veteran, as a back-up pilot, and both are now in Utah for a full dress rehearsal. The Utah range is as much a stranger to NASA as the pilots are — more used to chemical and biological warfare testing than landings from outer space. Its thousands of square kilometres of salt flats, marked by just a few dirt roads, were once considered an alternative landing site for the space shuttle. And NASA's Stardust project, now on its way to collect samples from the tail of a comet, will land here in 2007. But Genesis will be its first brush with space.

Down with a bang

A possible downside of the Utah site is revealed to reporters during their safety briefing: we are warned that unexploded ordnance still litters the base. Could Genesis land on one of these? "There is always a possibility," admits Clyde 'Rex' Rexroad, one of the mission controllers at the Hill Air Force Base down the road. But it's a very slim one, he assures us.

Today, with the capsule plummeting at 4 metres per second, Rudert swoops in for the catch. When the 6-metre pole and hook hanging from the back of his chopper make contact, a small piston drives a nail through the parafoil to hold it tight. Metres of line reel out to stop the load from feeling a shock, and Rudert slowly puts on the brakes. The parafoil now hangs limp, twisting in the wind. Fifteen minutes of slow, controlled descent puts the capsule gently on the ground. "I think he can fly that right into my pocket," says Don Sevilla, chief engineer for the payload at the Jet Propulsion Laboratory in Pasadena, California.

"It's just like fishing, only the fish doesn't fight back," Sevilla quips. He is only partly right. In tests the week before my visit, the team found that hooking the parafoil smack in its centre could have disastrous consequences. The chute can re-inflate, looking like a bow-tie instead of a simple ribbon, and fly back up towards the helicopter, forcing the crew to cut the line. To prevent this, the pilots now try to snag the chute on the side, leaving it flapping harmlessly.

Once grounded, the capsule has to be rushed to a temporary clean room on the base to prevent terrestrial particles and gases from contaminating its cargo. So from the moment Rudert makes his catch, the clock is ticking. "It's the most contamination-

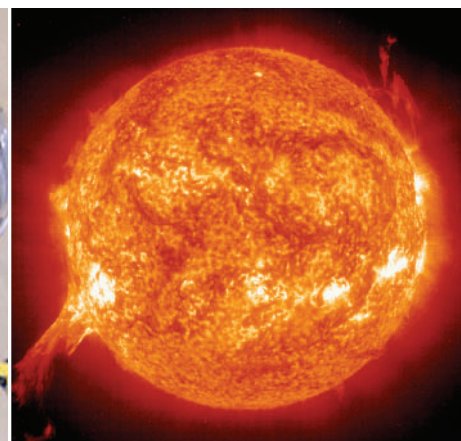
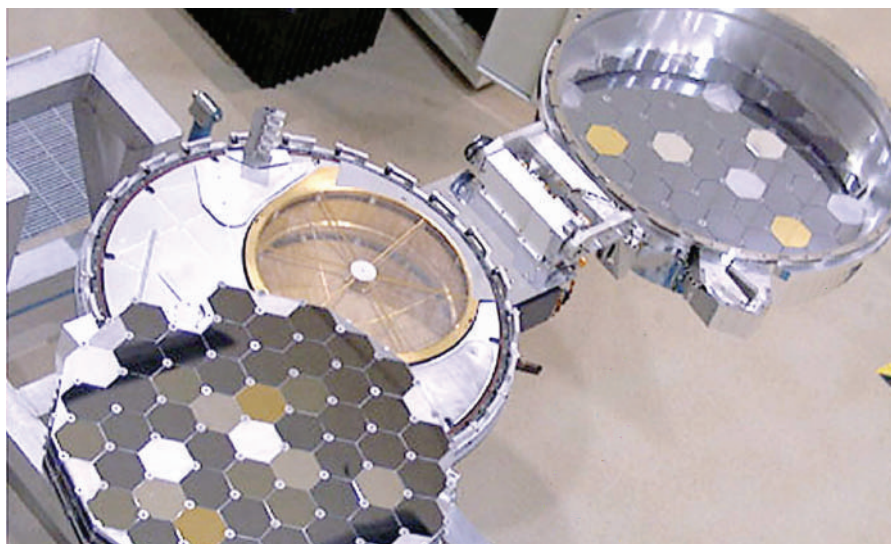
assured. Genesis's instruments would not survive a crash landing. The capsule carries a can full of delicate wafers of silicon, diamond, gold and sapphire, which basked in the Sun's rays for two-and-a-half years before heading home on 1 April. Smash this jewel box into the ground and scientists will spend years picking up the pieces.

Hence the need for a cushioned landing. A water landing would contaminate the capsule. Inflatable crash pads wouldn't be soft enough. The team considered a parachute landing for a while, until tests throwing the capsule off the back of a truck showed that even if it landed gently it could bounce and break up on the second impact. Luckily, NASA had another option.

Mid-air catches using planes have a long and successful history, starting with mail pick-ups in remote US locations in the 1930s. Planes swooped past a loop of rope suspended from two poles and snagged it, lifting the mailbag and winching it on board. In the Second World War, the same basic hook-and-winch system was used to pick up troop-carrying gliders from the ground. Similar technology was used to grab the parachutes of film canisters dropped from the first spy

"Most pilots sit up and look out the front. But we have to fly with our heads stuck out the window, looking down."

— Cliff Fleming



Catching the rays: the array of fragile wafers (bottom left) in Genesis (top left) have spent the past couple of years collecting particles from the solar wind.

that will stop us getting to it is a bad break in the weather," he says.

High winds, rain, poor visibility and low cloud cover — all these could scupper a catch. September usually means good weather in Utah, unless it is struck by a southwest monsoon or an Alaskan cold front. On 8 September 2002, there was a Pacific hurricane, notes Burnett. "8 September 2003 wasn't nice either," he says.

Still, on two practice days in April, a sandstorm, cold winds and a freak snowstorm didn't disrupt the plans (although Rudert did have to go shopping for extra-long underwear). If the weather looks really bad, Genesis can be put in a holding orbit for six months, returning to Utah in March 2005.

If, on the other hand, the parafoil doesn't open at all, there will be little hope. "We'll have a mess. A big mess," says Burnett. A duplicate parafoil, packed at the same time as the one on Genesis, was opened last week without incident. But this one wasn't exposed to the temperature extremes and radiation that have blasted the real thing. Sevilla reckons it is strong enough, although he worries about it being in space for three years.

A lot is resting on the shoulders of Fleming and Rudert, but they seem to be taking it in their stride. In the end, they say, it is no more stressful than pulling skiers from the bottom of crevasses, depositing fire-fighters in the midst of burning woods, or re-enacting the Vietnam war for the movies — all tasks that adorn their CVs. "No. I'm not nervous. Yet. When you think it is worth millions and has come from outer space, then you go 'woah,'" says Rudert. "There are a lot of things that have to happen before we even get involved," adds Fleming coolly. "If that doesn't go like clockwork we might as well go back to the bar."

Nicola Jones is *Nature's* assistant news and features editor in London.

♦ genesismission.jpl.nasa.gov

♦ www.moviepilots.com

sensitive mission ever flown," says Corwin. He picks up a few grains of sand from the desert floor. "We're bringing back less than that," he says. Today, the team makes the trip in under two hours — well within the mission's goal.

Once secured, the Genesis capsule will be ferried off to another clean room at the Johnson Space Center in Houston, Texas, for archiving and study. We have some hints about what they will find. Back in the 1970s, the Apollo missions showed the value of collecting samples from the solar wind. Astronauts unfolded a strip of aluminium, left it out for 45 hours, rolled it up and brought it home. This proved that if you put equipment out there, the solar wind would stick to it. And it gave us some firm data: the proportion of neon isotopes in the Sun.

"That's the one solid number we have and it has had a huge impact," says Burnett. There are more of neon's lighter isotopes in the solar wind than there are on Earth — by about a third. Researchers think this means that Earth once lost much of its atmosphere, bleeding away the lighter atoms. To check, Genesis will look for the isotopic content of nitrogen in the Sun. If this, too, is lighter than it is on

Earth, this popular theory will be confirmed.

But the real prize will be the oxygen isotopes, for which theories of the early formation of the Solar System predict certain numbers. To measure those, ultra-sensitive mass spectrometers have been designed to cope with extremely small, reactive samples. "This is breaking new ground analytically," says Karen McNamara, who works at the Johnson Space Center archive and will supervise the deconstruction of Genesis.

Grains of truth

The most valuable sample will be found in an electromagnetic dish that was designed to bend the ionized particles in the solar wind, concentrating them onto a cookie-sized collector. Researchers are most keen to protect this dish from the dents of a crash landing, as its exact shape will help them to work out how many solar rays have been bent in and trapped. Damage it, and this information is gone for ever — it cannot be put back together like so many shards of broken wafer.

Burnett is optimistic that Genesis will make a safe landing. But it isn't a sure thing. "We'll find it. We'll catch it. The only thing

Call for action to protect free exchange of ideas

A US law that limits trade to embargoed countries is now affecting scientists' activities.

Sir— Scientists and engineers need to take action to minimize the collateral damage to science caused by the war on terrorism. At issue here are US prohibitions on providing goods and services to people in countries embargoed by the United States, including Iran, Burma, Sudan and Cuba (see “Publishers split over response to US trade embargo ruling” *Nature* 427, 663; 2004).

Reversing an earlier ruling, the US Treasury ruled on 2 April that the Institute of Electrical and Electronics Engineers (IEEE) could edit and publish papers submitted by Iranians without obtaining a special licence. However, the same ruling “would consider a prohibited exportation of services to occur when a collaborative interaction takes place between an author in a Sanctioned Country and one or more US scholars resulting in co-authorship or the equivalent thereof”.

The implications of these rulings are severe. In 2003, more than 200 scientific articles were jointly written by authors with both Iranian and US addresses. A dozen or more English-language journals

are published in Iran, and a few Americans sit on their editorial boards, occasionally publish articles in them, and often are sent manuscripts by their editors for review. At least one US university has advised its faculty not to review manuscripts sent to them by editors of Iranian journals. Scientists are vulnerable, at least theoretically, to large fines for violating these embargo laws. The IEEE continues to deny all member services to people in embargoed countries, including awards and advancement to fellow status.

While there is no indication that the US Treasury is targeting the scientific community, at least one scientific society has already run into trouble: the International Union of Pure and Applied Chemistry was fined for an illegal currency transaction that involved a small expenses payment to a Russian scientist, who unexpectedly cleared the cheque through an embargoed Russian university.

These rulings raise important issues for US scientists and their colleagues around the world. As the Association of American University Presses pointed out in a press

release on 5 April, these rulings represent bureaucratic overreaching by the government. They provide no conceivable benefit to international security. Indeed, scientists and engineers in Iran and other embargoed countries are just the sort of people to whom Western democracies should reach out. They are well educated, often familiar with Western institutions, and technologically literate. They often occupy responsible positions in their countries.

The rulings should be a call to action by scientific and professional societies, and indeed to all groups interested in the free exchange of ideas. The academic and scientific communities in the United States need to respond to these challenges, preferably coordinated by the National Academy of Sciences and the National Academy of Engineering.

The international scientific community is also affected, and should protest loudly against these restrictions.

Kenneth R. Foster

Department of Bioengineering, University of Pennsylvania, 120 Hayden Hall, 3320 Smith Walk, Philadelphia, Pennsylvania 19104-6392, USA

Efforts to help Africa are tripped up by red tape

Sir— Your Editorial “A fair deal for all” (*Nature* 428, 451; 2004) recommends that the United Nations Educational, Scientific and Cultural Organization (UNESCO) should expand its agreements to redistribute used instruments to scientists in poorer countries. Having run the Scientific Apparatus Recycling Scheme (SARS) for the Federation of European Biochemical Societies (FEBS) since 1991, and also being involved in the support of African biochemists by UNESCO, I fear that unless the administration of UNESCO grants can be greatly simplified, it will be increasingly difficult to find scientists with the appropriate experience to undertake this voluntary work.

Since 1999 I have received a series of small grants from UNESCO to support the provision of small quantities of biochemicals and spare parts to African scientists. The grants are paid to my university account so that the cost of the goods can be charged to this account by the suppliers, who then dispatch direct to Africa. This arrangement simplifies the task for the recipients and avoids multiple currency conversions, but involves many e-mails to ensure that the correct goods are

delivered. The main problem I have had is with the bureaucracy of UNESCO, which — although designed to ensure that I operate honestly — has little understanding of the problems at the coal face.

In the case of my most recent grant, I was told in December 2002 that I would receive US\$10,000 but, after much correspondence concerning the contract, which was issued in May, I received \$8,500 the following November and \$1,308 in January 2004, with no explanation of the shortfall. I had budgeted carefully against an exchange rate of \$1.6 to £1, but by January the dollar had fallen to 1.85. As a consequence of this and of the shortfall in funds, I was left £300 in debt. I informed UNESCO of this but have had no redress.

My wish is that UNESCO would operate like granting bodies in the United Kingdom: telling the recipient the grant available, promptly transferring the full funds and demanding a precise statement of the expenditure. If recipients do not perform correctly they know they will receive no further grants.

My experience with the SARS programme has been mostly positive. Since 1992, 107 loads of books, journals and apparatus have been sent to the FEBS societies in central and eastern Europe at a transport cost of more than £160,000

(US\$286,000). Most of the gifts have come from the United Kingdom but SARS also covers the cost of transport between countries on the European mainland. I think that if one started to charge for used instruments, as in some of the schemes mentioned in the Editorial, the administrative expenses would climb considerably.

Finally, your Editorial and later correspondence by Stevens Rehen *et al.* (“Scientific aid to Brazil is strangled by red tape” *Nature* 428, 601; 2004) noted customs barriers to scientific donations. I have had considerable difficulty with certain countries but often, if a nominal value is stated for each piece of equipment, the customs are satisfied. It also helps to get the load classified as humanitarian aid rather than technical aid, in order to avoid taxes, but this may take several months. A more positive attitude by governments to our problems would be welcome, but politicians often tell us “The customs are a law unto themselves”.

Peter N. Campbell

Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, UK

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

Nature's biological weapon

The 1918 flu pandemic killed 50 million people — and it could happen again.

The Great Influenza: The Epic Story of the 1918 Pandemic

by John M. Barry

Viking: 2004. 546 pp. \$29.95

John Oxford

In the modern world where a newly discovered virus has been called *sin nombre*, or 'nameless', to avoid pinpointing a geographical origin, a strange paradox has arisen — several nations are claiming Spanish influenza as their own. But virologists are all agreed on one thing: the virus did not emerge in Spain. A strong US contingent insists that it arose in army camps of 60,000 soldiers in Kansas and then spread to Europe via troop ships. My own opinion is orientated towards the European Union: there was an initial focus in British army camps in France and Britain in 1916 and 1917, and then a gradual seeding over the next 12 months before the virus burst upon an unsuspecting world. So shall we rename this killer as the greatest US bioterrorist weapon of all times, or English or French? Of course not — the virus will always be known as the Spanish flu or, even worse, the Spanish Lady.

Two superb recent papers (S. J. Gamblin *et al.* *Science* **303**, 1838–1842; 2004 and J. Stevens *et al.* **303**, 1866–1870; 2004), which detail the X-ray crystallography of the major surface antigen of the 1918 virus, are full of descriptions of 'Spanish' influenza. I believe that from the gas-ridden, overcrowded trenches and nearby hospitals of already ruined Europe, filled with enough pigs, chickens and ducks to feed 2 million troops each day, arose a 10,000-nucleotide beast surrounded by a fragile lipid sphere. It seeded itself around the world while its biological clock ticked to ring at midnight on 11 November 1918, when it exploded worldwide.

In a world already shattered by death from high explosives, what more could possibly happen? The deaths of 50 million more innocents, from Spanish flu. The pictorial image is not far removed from Pieter Bruegel the Elder's *Triumph of Death*, with bubonic plague and the pale rider on a pale horse gathering in souls at random. I had not fully appreciated US President Woodrow Wilson's demand of the American nation once the decision had been made to go to war: "the spirit of ruthless brutality and deathless determination of total war". This almost totalitarian command led directly to an infectious-disease tragedy in the United States when the virus broke out in army camps and on troop ships to Europe. Wilson's



Death rides again: the 1918 flu pandemic caused scenes like those in Bruegel's *The Triumph of Death*.

government knew the risks but forced the troop movement onwards nevertheless.

This is the setting for John Barry's *The Great Influenza*. Barry is a philosopher who writes like an angel. Through a vision of the scientists of the day — Flexner, Welch and Avery — he brings back the tension and excitement, the despair and the sorrow. The only previous book that captured this was *They Came Like Swallows*, a 1937 book in which William Maxwell quietly and beautifully reconstructs his Edwardian childhood in Connecticut. His mother died of the flu and he was changed forever, never again feeling safe. But Barry's writing matches this while still managing to capture the science of virology. We read about viral swarms and quasi-species, passage from human to human to produce virulent viruses, reassortment and genetic recombination. A cytokine storm is so well described that I will borrow the phraseology to explain it to my own students. But this is not a textbook of influenza — that is its strength. The author knows his ground well and stays there.

Barry has made some serious decisions about the form and nature of his book and I, with one exception, agree wholeheartedly with him. There are no modern scientists

here but our predecessors by two or three generations — virology is just emerging, from a world of chocolate agar plates, Bunsen burners and the Gram test. My single grouse is that book is almost entirely focused on the United States. A first reading so swept me along with the narrative that I overlooked this omission. But on a second look — and yes I have indulged myself and read the book twice — the uni-nationalist flavour slightly impedes the scientific story. Barry pulls information from everywhere, from recent TV films, newspapers of the times and scientific journals, and, to the best of my knowledge, it is accurate as it can be.

One of the virus's victims was William Osler, a member of the gang of four who founded Johns Hopkins School of Medicine in Baltimore. I never had much sympathy for Osler and his textbook, with its quotes about "pneumonia being the old man's friend". Here we read how he caught flu as a 70-year-old. In November 1919, in another wave of infection, he had "paroxysms of coughing", his pleural attachments were ripped to smithereens, and nearly a pint of pus was drained from lungs. For three months he struggled. He quoted Tennyson "of happy men that have the power to die",

THE ART ARCHIVE/MUSEO DEL PRADO, MADRID/DAGLI ORTI

and on 29 December his wish was granted. He died of virus-induced pneumonia. How can this disease be anyone's friend? But this description of bronchopneumonia must have been more typical of the deaths in 1918 than the exaggerated tales of blood and gore. It was bad enough in its quietness, strangulation and breathlessness, and needs no exaggeration.

These post-1918 epidemic waves of flu last until 1922 or so, and then we encounter, albeit briefly, the outbreak of 'sleepy sickness', which immobilized another 5 million souls around the world.

Through page after page we relive that period and can ask ourselves whether it could happen again. Of course it could, and we are no better prepared to resist it. Influenza virologists know this, the World Health Organization knows it and so do public-health specialists. But still we seem to be suspended as if in a dream, like Alice in Wonderland. There is fine talk of vaccines and antivirals, but if such a pandemic influenza A virus were to jump the species barrier this month, we could all be caught in a catastrophe as serious as that of 1918.

This book is a wake-up call. It should reach beyond our small, incestuous coterie of a few hundred flu virologists into a far greater world that will not like what it sees. The theme here is that a new virus will come. We have the knowledge and the scientific means to blunt its virulence, but the attention of politicians and decision-makers has been diverted towards bioterrorism and war.

Should this timely book prove a turning point, then we could label Barry as a saviour, although it is clear that this was not his intention when writing the book. He totally immerses us in the 1918 flu battle. These first scientists challenged the natural order, and when the great influenza came they placed their lives in the path of the disease. The scientific knowledge that is still coming out of studies of the 1918 pandemic points us "to much that lies in medicine's future". The ordered, quiet world of Maxwell has vanished. But no amount of slick presentation or clever lawyers' talk from our leaders will escape this volcano sitting in all our countries.

Influenza was the twentieth century's weapon of mass destruction, killing more than the Nazis, more than the atomic bomb, and more than the First World War. We would do well to dwell very seriously indeed on this fact. Nature is the greatest bioterrorist of our world, and we should concentrate and expand our efforts in public health. Emerging viruses could do for us all, as easily and as quickly, or even more so, than the Great Influenza of 1918. ■

John Oxford is at Retroscreen Virology Limited, Centre for Infectious Diseases, Institute of Cell and Molecular Science, Queen Mary's School of Medicine and Dentistry, University of London, London E1 4NS, UK.



STEPHEN FRINK COLLECTION/ALAMY

Deep trouble: turtles, once a common sight on Caribbean reefs, have been destroyed by hunting.

Taking stock of conservation

Against Extinction: The Story of Conservation

by William M. Adams

Earthscan: 2004. 328 pp. £55 (hbk), £16.95 (pbk)

E. J. Milner-Gulland

I recently observed a UK delegate at a conference lecturing a representative of a developing country's wildlife ministry on how they should be doing conservation. The question under discussion was not the one that might have been asked a few decades ago — "What are you doing for the species?" — but rather, "What are you doing about rural poverty?"

Bill Adams' book *Against Extinction* encourages us to reflect on the continuities and upheavals in the way we do conservation. As Adams insightfully points out, most conservationists have three time-frames: the long term, for comparing extinction rates, for example; their lifetime, for nostalgia about paradise lost; and the immediate crisis, for most of their interventions. He wants us to look at conservation on another time-frame — the past century — to help us understand the route that led to our current conservation philosophies, and, in so doing, to reflect on our unwitting prejudices and assumptions.

Ignorance of recent history, described by Daniel Pauly as the "shifting baseline syndrome", is a pernicious and prevalent problem in conservation. J. B. C. Jackson has pointed out that current observations of Caribbean reefs are of an ecosystem that has already been fundamentally altered from its natural state by human intervention. Observers in the fifteenth century, before turtle populations were ravaged by hunting, described densities so high that it seemed ships might run aground on them. Jackson memorably described present-day Caribbean reef systems, which have few turtles, as being like the Serengeti without the ungulates. There are many examples in which our ignorance of human and ecological history clouds our understanding of contemporary events. So I applaud Adams' insistence that we look back to obtain a true perspective on the conservation movement today.

Adams' definition of conservation is narrow, and he makes it clear from the start that he is considering only the birth of conservation among Western colonial nations, starting around the middle of the nineteenth century. Central to the book is an organization now called Fauna and Flora International, which celebrated its centenary last year. It has led the way in Western conservation thinking throughout the past century, changing from its origins as an influential group of colonial hunters primarily focusing on game animals in Africa to its current role, working through local, non-governmental

organizations and concentrating on in-country capacity-building.

I can sympathize with the need to tell a manageable and coherent story, but it is frustrating to have such a narrow view of the roots of conservation perpetuated, particularly by someone who is well aware of this view's limitations. To take just one example, the Mongolians pride themselves on having set up the world's first National Park in 1778 — as do the Americans in 1872. Clearly there are issues surrounding the definition, but it is a shame that the latter view of history is ubiquitously and uncritically repeated in conservation texts.

One of the pleasures of the book is Adams' explanation of the complex intertwining over the past century of the concepts of sustainable use, hunting and wildlife preservation. Some concepts, such as biodiversity and sustainable development, are recent additions; others have shifted in and out of fashion, metamorphosing as they went. Although the early conservationists tended to be exclusionary in their outlook, they were well aware of the potential of sustainable use as a conservation tool. We can be sure that most of the 'new ideas' of conservation were actually reawakened as the dynamic culture of conservation shifted again.

Most of the book is determinedly factual and full of detail, with many interesting examples and case studies — although readers will be hard-pressed to find them again, as the book has no bibliography or visual aids such as chronologies, relying instead on chapter-based endnotes and an inadequate index. The book will also be frustrating for those who wish to find scientifically based analyses of the pros and cons of different conservation approaches; that is not the book's aim. The final chapter, in which Adams outlines his vision for conservation, seems to belong to a different book. The ideas in it are challenging and raise fascinating questions, but seem strangely disengaged from his previous careful historical analysis.

One of the things that makes conservation interesting and challenging is that its practitioners have many perspectives and disciplinary backgrounds. This cultural diversity is growing as social scientists become more fully engaged. One of the great hopes for conservation in the future is this need to look at issues through others' eyes. Just as ecologists now need to be able to perform quantitative analyses, so recently trained conservationists cannot get by without learning social science. This book is a major contribution towards opening conservationists' eyes to another world of historical and cultural understanding, which I welcome wholeheartedly. ■

E. J. Milner-Gulland is in the Department of Environmental Science and Technology, Imperial College London, Exhibition Road, London SW7 2AZ, UK.



All at sea: climate change is looming over us, but will it really leave New York under water?

Film

Making heavy weather

The Day After Tomorrow

Directed by Roland Emmerich

20th Century Fox

Worldwide release on 28 May 2004

Myles Allen

I have yet to meet a doctor who doesn't dismiss the TV drama *ER* as hopelessly unrealistic, and yet who doesn't tape it religiously if they happen to be on call. I've also yet to meet a doctor who doesn't regard meteorologists and oceanographers as spotty geeks who couldn't possibly be doing anything glamorous enough to be worth a TV series, never mind a blockbuster Hollywood film. So, with the release of *The Day After Tomorrow*, a blockbuster-and-a-half inspired by the issue of human-induced sudden climate change, we must be careful not to confirm the medics' worst suspicions by pedantically carping on about the film's portrayal of geophysical fluid dynamics.

A medic watching this film would learn as much about climate as I would learn about cardiology watching *ER* — not nothing, but I would prefer the surgeon standing over me with a scalpel, or the politician pondering my petrol taxes, to have had some additional training. So I find the fuss about the film's possible impact on climate policy rather disturbing. Bjørn Lomborg vehemently attacked the film recently in the *Independent on Sunday* for bouncing politicians into signing the Kyoto Protocol. It's a film, lighten up. I'm sure the world's teenagers can work out that this is hardly exam revision material, and if it inspires a few of them to stick with physics for a couple more years and perhaps consider a university course in the geosciences, then it will have more than justified its special-effects budget.

Could *The Day After Tomorrow* do for

meteorology and oceanography what *Top Gun* did for US Air Force recruitment? The special effects are stunning and the filmmakers have clearly gone to some lengths to base them all on natural phenomena, although the connections between them are more tenuous. A tidal wave could indeed hit New York, albeit one more likely induced by a submarine landslip than a gigantic storm surge. Strange things do happen in the eyes of hurricanes, although to get stratospheric temperatures at sea level you have to be fairly creative with your thermodynamics. I draw the line at someone embedding a hurricane model into a global weather model in 48 hours, but perhaps it is wise not to tell the teenagers what climate modelling actually involves until after they have signed up.

I believe that the public takes a much more sophisticated line than Lomborg fears. I am involved in a public-participation experiment (www.climateprediction.net) that is looking, among other things, at how the atmosphere might reinforce a thermohaline slow-down. Contributions from the public on the discussion boards have generally been level-headed. Everyone understands that there's a link to issues raised by the film without mistaking the film for a forecast.

So, the film is well worth a lab night out, particularly if your model is giving trouble. Perhaps the hardest part will be judging how to respond to questions in the pub afterwards about whether this has anything to do with our actual projections for human-induced climate change. We have to be clear that the film is science fiction, but we also have to make sure we don't belittle what is actually going on. A prescient dinosaur, gazing future-wards over the millennial undulations of global temperatures, would probably just about make out the warming spike representing our humble contribution to the twenty-first century. It's quite an ego-boost, isn't it? The last species to have this much influence on the climate was almost certainly green, slimy and inarticulate. A teenager signing up for the geosciences today

is guaranteed an interesting career, for while unfettered anthropogenic climate change will certainly not turn out exactly like *The Day After Tomorrow*, it should still be a show worth watching — after *ER*, of course. ■

Myles Allen is in the Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK.

The second creation

Human-built World: How to Think about Technology and Culture

by Thomas P. Hughes

University of Chicago Press: 2004. 224 pp. \$22.50, £16

Graham Farmelo

Five years ago, the US Army awarded a \$45-million contract to the University of Southern California to establish the Institute for Creative Technologies. The funding of this organization — a confluence of the interests of the military, academia and Hollywood — was a significant step in the history of American technology, from President Eisenhower's 'military-industrial complex' towards something quintessentially modern, a military-entertainment complex.

This development has not evoked much comment, probably because, as the distinguished historian Thomas Hughes laments, most people in the industrialized world have a simplistic view of technology. They think of it as merely being about computers and other gadgets, and as "a handmaiden of commercial capitalism and the military". Hughes seeks to put this right in *Human-built World*, in which he presents an accessible, multi-disciplinary review of modern technology.

Hughes concentrates on the development of technology in the United States, nodding only occasionally to Germany. In the first and best chapter, he describes how the European settlers viewed their transformation of the wilderness into a purpose-built environment as "a second creation". Technology was a gift from God. This religious perspective led the historian Lynn White Jr to note that the book of Genesis in the Bible has persuaded many Americans that God gave them a privileged dominion over nature, an observation that sheds a good deal of light on the environmental policies of the present US administration.

The deployment of new technology in the United States has not been without its critics. After generations of industrialists had enthusiastically followed Henry Ford's precepts for mass production ("system, system and more system"), a backlash was inevitable. Hughes vividly reports the opposition (which reached a peak during the Vietnam

Exhibition

Scientific expressionism

In the mid-nineteenth century, Guillaume Duchenne, who described the form of muscular dystrophy that bears his name, documented in detail all the muscles of the human face and the facial expressions that they could convey.

Fear, joy, horror, disdain and disgust — Duchenne could reproduce any expression of emotion by direct electrical stimulation of the appropriate group of muscles. In doing so, and in recording his experiments photographically, he caused a sensation. At the time, our ability to convey subtle emotions on our faces was considered a divine manifestation of the inner consciousness that separated us from the beasts — not merely a matter of simple physiology.

An introverted man and an unconventional scientist, Duchenne was not shy of challenging the world. Using this series of facial shots, he argued that the ancient Greeks often got it wrong.

The facial expressions on their sculptures did not always accurately reflect the anatomy of the emotions that the artists intended to convey, he claimed.

Duchenne was one of the first scientists to use the new technology of photography as part of the scientific process. The image shown here is included in the exhibition "Photography and Painting in the Nineteenth Century", which runs until 18 July at the Kunsthalle der Hypo-Kulturstiftung in Munich, Germany.



The extensive exhibition describes how painters were influenced by the new way of seeing, and how scientists, engineers and architects used photography as a means of record-keeping — which also allowed them to see their own worlds differently. It emphasizes the unexpected dialogue between science and art, which were both confronted with this revolutionary new tool at the same time.

Alison Abbott

♦ <http://www.hypokunsthalle.de/newweb/index.html>

War) to the systems approach by humanists, public intellectuals and artists, who pointed to what they saw as the consequent erosions of personal freedom and deterioration of the environment.

In the light of this history of opposition, Hughes seems surprised by the wide-eyed reception that the public has given over the past two decades to the burgeoning of digital information technology. Even this technology has religious connections, he points out: George Gilder, the high-profile celebrant of the information age, has long highlighted the emerging 'revolution' in information technology as a kind of new religion.

We do not need to share Gilder's prophetic zeal to agree that information technology is the nearest thing we have seen in contemporary life to a technological revolution. For good or ill, it has substantially changed the way that huge numbers of people live in the industrialized world. Yet Hughes is sceptical about this, and dismisses the claims of its most passionate advocates, notably Nicholas Negroponte. The visionary contributions of Marshall McLuhan go unremarked.

Hughes' passion obviously lies elsewhere, in projects that seek to harmonize techno-

logical developments with the environment — the field of ecotechnological environmentalism (never has a human endeavour been in greater need of a new name). His conclusion, a lengthy panegyric to Florida's project to restore its Kissimmee River system, underlines his optimism. But one wonders what impact this project will have, compared with the environmental destruction that developers are currently inflicting on other parts of the state.

Human-built World is a rewarding if unsatisfying book, too dense to appeal to lay readers but too light to be of much use to scholars. It is, however, a virtuoso overview of the various relationships between technology, commerce, society, art and the military. To anyone who has read it, the emergence of a military-entertainment complex will appear as natural as it did to Milo Minderbinder in *Catch 22*: "Frankly, I'd like to see the government get out of war altogether and leave the whole field to private industry." One wonders whether similar thoughts have crossed the mind of Donald Rumsfeld in the past few weeks. ■

Graham Farmelo is senior research fellow at the Science Museum, London SW7 2DD, UK.

A late change to the programme

How an engineer became hooked on atmospheric chemistry.

Paul J. Crutzen

As a child, I dreamed of being a scientist, but when I left middle school this looked impossible. Because of illness during my final exams, my grades were too low to obtain a stipend to continue to university, and so I switched to a shorter course in civil engineering. I found it interesting — in my first job as an engineer, I helped to construct a pedestrian tunnel between Amsterdam's central railway and the tram stations. Unfortunately, few used it, and so it had a short lifetime. In 1958, I married a Finn and moved to Sweden, where I worked for a construction company. But I had not given up hope of higher learning.

My chance came when the Meteorological Institute of Stockholm University advertised for a computer programmer. There were many applicants and I did not have any experience, but I got the job. So in the summer of 1959, with a baby daughter in tow, my wife and I moved to Stockholm, where I was to learn a new profession — computer programming. At last, at the age of 26, in between writing programs for various meteorological applications, I could attend some lectures at the university. I had to concentrate on theoretical subjects, so I studied mathematics, mathematical statistics and meteorology in my free time.

By the middle of the 1960s I was helping a visiting scientist, James Blankenship, with calculations on the vertical distribution of atomic oxygen and ozone in the upper atmosphere. We were using Sidney Chapman's five 'oxygen-only' chemical reactions, first described in 1930, for how sunlight converts the various forms of oxygen — O , O_2 and O_3 — into one another.

I became hooked on stratospheric ozone chemistry. I avidly began to study the literature, and became fascinated by catalysis, which is so characteristic of atmospheric chemistry. Some studies at that time had suggested that OH and HO_2 radicals control stratospheric ozone. In my *Filosofie Licentiat* thesis (comparable to a PhD) I pointed out that the rate coefficients assumed in those studies — they had not been experimentally measured — could not explain the vertical distribution of ozone and would have led to unrealistically rapid ozone destruction in the troposphere. Soon after, laboratory investigations showed that the reaction coefficients were indeed much too high.



The author (centre) with his *Filosofie Doctor* examiners, Richard Wayne (left) and John Houghton, at Stockholm University.

In 1969, I was awarded a stipend from the European Space Research Organization to join John Houghton's research group at Oxford University as a postdoc. Houghton had developed techniques to measure trace gases in the stratosphere from satellites, and soon after my arrival in Oxford, he showed me optical spectra taken in the stratosphere by David Murcray and colleagues from the University of Denver, Colorado. Intriguingly, these spectra contained features of HNO_3 . I reasoned that if HNO_3 was in the stratosphere, then there must also be NO_2 , as this is produced by photodissociation of HNO_3 by solar ultraviolet radiation.

In 1970, encouraged by these findings, I proposed a simple reaction chain, involving NO and NO_2 as catalysts, that destroys ozone in the stratosphere. Because NO is produced by oxidation of N_2O , which in turn originates from microbial transformations at the Earth's surface, this demonstrated that biological activity strongly influences stratospheric ozone chemistry. In the following year, Harold Johnston and I both called attention to the possibility that the NO emissions from large fleets of high-altitude, supersonic aircraft (that were to be built in the United States, the Soviet Union, Britain and France) would damage the levels of stratospheric ozone. Major international research programmes were started to determine more about the physics and chemistry of the stratosphere — which at that time was also dubbed the 'ignorosphere' as so little was known about that

region of the atmosphere. These studies confirmed that at altitudes above roughly 20 km, the NO emissions could indeed deplete ozone levels, but also showed that the original ozone depletion estimates were too high, because the effect of some producing reactions, similar to those that produce ozone in the atmosphere, had been underestimated. Those were nervous years, as I was proposing new atmospheric chemistry without a proper background in chemistry. However, a year spent at Oxford University's Physical Chemistry Laboratory with Richard Wayne helped me immensely.

As it turns out, NO emissions are growing only rather slowly, as the planned fleets of supersonic aircraft did not materialize. The greatest threat to ozone levels was revealed in 1974 by Mario Molina and Sherwood Rowland, who showed that Cl and ClO — photolysis products from industrial chlorofluorocarbons — strongly deplete ozone. In 1985, Joe Farman and colleagues surprised the scientific world when they reported the springtime 'ozone hole' over Antarctica, a place where ozone had always been supposed to be inert. But that is another story.

Despite my late start in academia, I had the great fortune to enter a research field in which there was so much to discover. To describe the chemistry of the atmosphere, we now need hundreds of reactions, many more than Chapman's five reactions. Much knowledge has been gained in atmospheric chemistry since the early 1970s — for example, air pollution does not only come from fossil-fuel burning, but also from extensive tropical biomass burning; and in extreme scenarios, burning cities and industries during a nuclear war could cause 'nuclear winter'. Most research in atmospheric chemistry has to do with humankind's strong interference with nature — we live in the 'anthropocene'. Fortunately, an international agreement has been reached to ban the production of ozone-depleting substances — an important turning point in global environmental policy. Although it is hard to be optimistic, let us hope that agreements on global climate and nuclear disarmament will follow. ■

Paul J. Crutzen is at the Max Planck Institute for Chemistry, 55020 Mainz, Germany, and at the Scripps Institution of Oceanography, Center for Atmospheric Sciences, University of San Diego, La Jolla, California 92093-0221, USA.

Automata make antisense

Anne Condon

Information-carrying DNA strands can be used to perform simple computations, but have so far been little more than toys. Can molecular computers be more broadly useful — in medicine, for instance?

People have long been fascinated with automata, fashioning them from any available materials to create mechanical creatures or musical devices, or to carry out simple computations. The materials used have even included biological molecules: in 2001, for instance, Yaakov Benenson and colleagues¹ built a tiny automaton from DNA strands and enzymes. On page 423 of this issue², Benenson *et al.* suggest how such molecular automata might be used to augment so-called antisense technologies, carrying out a diagnosis *in vivo* (that is, in a living cell) that automatically controls drug delivery.

Antisense drugs are oligonucleotides — short, single-stranded DNA molecules — that offer the promise of treating diseases caused by the expression of a harmful gene, for example a cancer-causing gene³. These drugs are designed to prevent the expression of such a gene: they bind specifically to the messenger RNA (mRNA) strand that is transcribed from the gene, thereby inhibiting the translation of the mRNA into a protein.

Benenson *et al.*² were motivated by the idea of achieving the 'conditional' release of such an antisense drug: *if* certain diagnostic conditions are true, such as low expression levels of certain mRNAs and high expression levels of others, *then* the drug is released. The *if-then* mechanism is the new element of computation introduced in their scheme.

Benenson and colleagues' diagnosis proceeds in specific steps (transitions), one for each of the diagnostic conditions. After each transition, the state of the diagnosis is either positive ('yes'), indicating that all conditions tested to date are true, or negative ('no'), indicating that at least one of the conditions tested so far is false. The diagnosis can thus be viewed as a sequence of transitions of a molecular automaton that can exist in two types of state (Fig. 1). Each transition tests for high or low levels of a particular indicator molecule.

In Benenson and colleagues' experiment, which was done *in vitro* (that is, in an artificial environment rather than in a cell), the drug is enclosed in the loop of a hairpin-shaped oligonucleotide structure called the diagnostic molecule. Initially, the stem of the diagnostic molecule is composed of four guards, one for each of four diagnostic conditions that the authors test (Fig. 2a). A positive diagnosis for each condition results

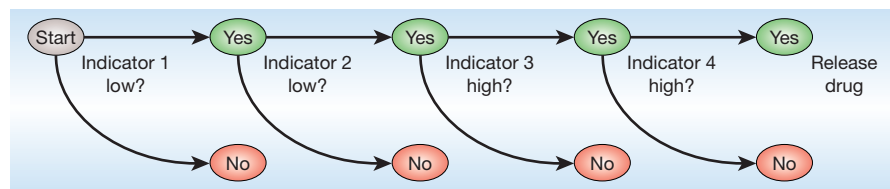


Figure 1 Automaton in abstract. Benenson *et al.*² devised a molecular automaton that theoretically tests for certain diagnostic conditions (high or low concentrations of particular indicator molecules) in cells. After each transition, the state of the diagnosis is either positive ('yes'), indicating that all conditions tested to date are true, or negative ('no'), signifying that at least one of the conditions tested so far is false. Should all four conditions be satisfied, indicating a positive diagnosis overall, a drug would be released.

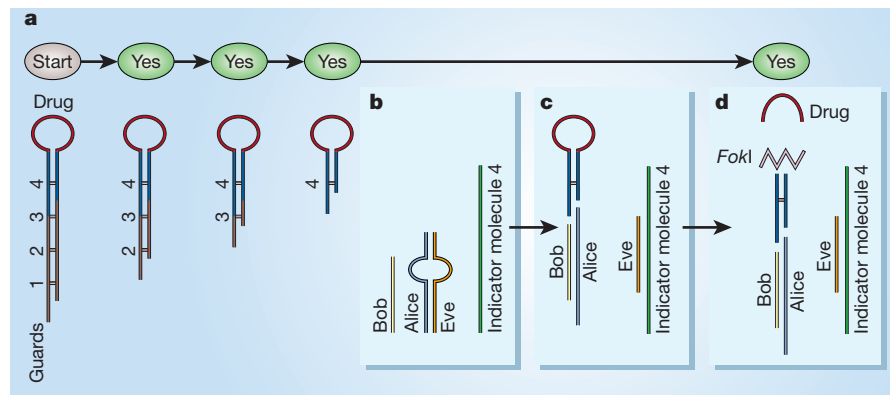


Figure 2 Benenson and colleagues' molecular diagnostic automaton. a, The start state of the automaton is a hairpin structure. The drug is a nucleotide sequence within this structure that is protected by four 'guard' sequences. Upon each 'yes → yes' transition, indicating low or high levels of an indicator molecule as appropriate, a guard is cleaved from the stem of the hairpin, yielding a new diagnostic state. b, Initial configurations of the members of the cast of oligonucleotides needed for the final 'yes → yes' transition. c, To effect the transition, allegiances shift among the oligonucleotides: Eve binds to indicator molecule 4, leaving Alice to bind to Bob. The Alice–Bob combination then binds to guard 4. d, The *FokI* enzyme (jagged line) recognizes the sequence formed when the Alice–Bob pair binds to guard 4, and then cleaves guard 4, producing the final 'yes' state in which the drug is released.

in removal of the guards, one at a time. The changing diagnostic molecule, which gets shorter with each transition as guards are sequentially removed, is the physical realization of the state of the automaton as the diagnosis progresses.

Physically, each transition is a carefully orchestrated game of shifting allegiances among several oligonucleotide 'players'. Consider the final 'yes → yes' transition, which should diagnose a high level of indicator molecule 4. In addition to the diagnostic and indicator molecules, the cast of players includes a pair that we will call Alice and Eve, and a lone ranger, Bob (Fig. 2b). Eve prefers to pair with indicator 4 rather than with Alice. So if all of the players are added to a

solution in which indicator 4 is present, then Eve deserts Alice and pairs with indicator 4. As a result, Alice settles for Bob (Fig. 2c). Furthermore, the Alice–Bob pair can dislodge the final guard, thereby releasing the drug (Fig. 2d). In the absence of indicator 4 in the solution, Alice and Eve would remain paired and the drug would remain captured.

How can the oligonucleotides be designed so that they are likely to shift allegiances according to this plot? Pairing, or formation of a duplex, between two oligonucleotides occurs when a sequence of nucleotides in one oligonucleotide binds to a complementary sequence in the other. Very roughly, the greater the number of nucleotide bonds between the two oligonucleotides,



100 YEARS AGO

Messrs. D. Schulte and Co. have submitted a sample of their self-lighting Bunsen burner, in which the well known property of finely divided platinum igniting under the influence of a stream of hydrogen is employed. The burner proper is of the usual type, but is furnished with a bypass tube at the side, controlled by a cross stopcock. At the top of the bypass, close to the open end of the burner, there is fitted a small bracket holding the bundle of several fine platinum filaments, so constructed that the thin stream of gas from the bypass tube impinges on the stretched wires... The arrangement works very readily, and if the old difficulties with regard to the durability of the delicate portions can be surmounted, the apparatus should be of considerable convenience to laboratory workers.

ALSO

At a sale recently held by Mr. Stevens in King Street, Covent Garden, a great auk's egg in fine condition was sold for two hundred guineas... This is a considerable falling-off from the three hundred guineas obtained for the last specimen sold by Mr. Stevens, the reason being attributed to the fact that several other fine examples are in the market. From *Nature* 26 May 1904.

50 YEARS AGO

It has been shown that as a rule the deoxyribonucleic acid content of the interphasic nuclei of tissues of a given species corresponds to a constant equilibrium value, which is double that of the deoxyribonucleic acid content of the spermatogenic nuclei of the same species. Consequently, at each mitotic division, synthesis of deoxyribonucleic acid must take place in order that the quantity of this substance should be restored in the nuclei of the daughter cells at a 'normal' level... According to some authors, this occurs immediately before mitosis, so that the content reaches double that of the normal value: after division each nucleus of the daughter cells receives also a normal content. According to other authors, synthesis occurs soon after mitotic division when nuclei of the daughter cells which receive only half of the normal content restore the latter. We have undertaken the study of this question in the thyroid cell of the white rat... These measurements show clearly that in our material deoxyribonucleic acid is synthesized immediately before the onset of mitosis. From *Nature* 29 May 1954.

the more stable the duplex. Eve is designed to be complementary to a relatively long unpaired region of indicator 4, whereas Alice is complementary only to a sub-sequence of Eve. So, Eve binds more stably with indicator 4 than with Alice. Also, the Alice and Bob oligonucleotides have (somewhat shorter) complementary sub-sequences, which are likely to form a duplex once Eve has bonded to indicator 4. The game of shifting allegiances plays out somewhat differently when diagnosing low, rather than high, concentrations of an indicator molecule, although the principles are similar.

Also needed is a mechanism for cleaving the guard in the presence of the Alice–Bob pair. This pair can bind to the guard (again, via bonding between complementary single-stranded regions of Alice and the guard, called sticky ends; Fig. 2c). One additional molecule, an enzyme called *FokI*, recognizes a sequence pattern formed when the Alice–Bob pair is bound to the guard. *FokI* then cleaves the guard from the hairpin loop, releasing the drug.

Unfortunately, the specific mechanisms proposed by Benenson *et al.*² would not work in a living cell: unwanted side effects of the

cast of supporting molecules (particularly the *FokI* enzyme) would be one major problem⁴. Nevertheless, getting an experiment of this scale to work *in vitro* is a real achievement. Perhaps more importantly, the work takes a conceptual step forward, by linking the development of molecular automata to antisense therapies. It is plausible that different molecular mechanisms can be found to create diagnostic automata in the cell, building on progress made so far in the use of antisense therapies or cellular computation^{5,6}. Developing such mechanisms would certainly be a good direction for further research.

Anne Condon is in the Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada.

e-mail: condon@cs.ubc.ca

1. Benenson, Y. *et al.* *Nature* **414**, 430–434 (2001).
2. Benenson, Y., Gil, B., Ben-Dor, U., Adar, R. & Shapiro, E. *Nature* **429**, 423–429 (2004).
3. Kurreck, J. *Eur. J. Biochem.* **270**, 1628–1644 (2003).
4. Stein, C. A. *J. Clin. Invest.* **108**, 641–644 (2001).
5. Weiss, R. & Knight, K. F. Jr in *Lecture Notes in Computer Science* 2054 (eds Condon, A. & Rozenberg, G.) 1–16 (Springer, New York, 2001).
6. Weiss, R. *et al.* *Nat. Comput.* **2**, 47–84 (2003).

Granular materials

The brazil nut effect — in reverse

Troy Shinbrot

In a box of mixed nuts, the brazils rise to the top. In granular mixtures in general, depending on their size and density, the 'brazil nuts' may sink instead. This reverse effect has now been explored further.

Every farmer can attest to the curious fact that the largest crop each spring is the boulders that appear, untended, on open fields. Common wisdom holds that this crop is loosened from the soil by frost heave, and rises because small pebbles can slip beneath large boulders, but not vice versa¹. This is the 'brazil nut effect' — named for the fact that, in a container of mixed nuts, the brazil nuts always seem to rise to the top (Fig. 1). Because similar processes and effects occur in pharmaceutical, chemical and food processing, the problem of granular segregation has earned serious attention² — and now, in *Physical Review Letters*, Huerta and Ruiz-Suárez³ add the latest piece of the puzzle.

The first complication to the simple picture of pebbles slipping beneath boulders (termed 'percolation') was the demonstration that a tapped bed of grains 'conveys' in a regular pattern: a wide swath of grains rises in the centre of a container, and thin margins correspondingly sink⁴. According to the convection picture, large 'intruder' particles rise with the surrounding bed, and then find themselves simply unable to fit into narrow

downwelling margins. This mechanism was confirmed by a clever experiment in which the convection rolls were reversed and, as predicted, large particles migrated to the bottom of vibrated beds⁴. Later confirmations came from magnetic-resonance-imaging experiments that conclusively demonstrated the presence of segregating convection rolls⁵, and from meticulous computational comparisons that revealed that convection dominates over percolation in producing segregation in deep beds⁶.

Over the past decade, however, our understanding of the segregation of large particles in vibrated beds has been challenged by experiments revealing that although large heavy 'intruder' particles can indeed rise in vibrated beds of finer grains, equally large light intruders can sink, contrary to expectation and common experience. Now termed the 'reverse brazil nut effect', this observation⁷, made by myself and Fernando Muzzio, is explained by neither the convection nor the percolation description. It is so counterintuitive that a reviewer of the original manuscript reporting the effect insisted that it could not be correct; and the



Figure 1 Nuts: why do brazils always rise to the top?

manuscript editor at *Physical Review Letters* conscientiously (if sceptically) tested the effect in his office using a jar of sand, a large plastic pin (a light intruder) and a steel nut (a heavy intruder). Since this impromptu confirmation, particle-dynamics simulations (subsequently validated experimentally⁸) have verified that the reverse brazil nut (RBN) effect appears under ideal *in silico* conditions, and that multiple intruders also separate in the curious RBN manner⁹.

In fact the RBN effect turned out to be even more complex than realized at first. Subsequent experiments showed that there are actually separate size and density influences at work in a tapped bed¹⁰. On the one hand, for intruders of a fixed density there is a distinct size threshold above which intruders rise, and below which they sink. On the other hand, intruders of a fixed size rise with a speed that grows and then diminishes non-monotonically as the intruder density is increased. To complicate matters still further, there have been numerous commentaries on the RBN effect, some of which question whether the effect even exists, or whether it is actually a computational artefact¹¹.

Huerta and Ruiz-Suárez³ have shown that there are actually two distinct regimes of segregation: one seen at higher frequencies of vibration (50 Hz), in which the bed becomes fluidized and ordinary buoyancy prevails

(heavy intruders sink but light ones float); the other at low frequencies (5 Hz), in which intruder inertia and bed convection conspire to produce either the ordinary or the reverse brazil nut effect, depending on intruder size and density. This result concurs with much earlier findings that there is a transition at vibration frequencies of about 20 Hz, below which dissimilar-size particles segregate and above which they mix¹².

Interestingly, this transition coincides with the frequency at which the surface first forms heaps driven by air flow (pumped by piston-like container motion against the granular bed), suggesting that the transition between ordinary buoyancy and the RBN effect is tied to air flow. This hypothesis has also been investigated recently, but the dual roles of grain and air dynamics remain entangled^{10,13}. Huerta and Ruiz-Suárez³ propose that this entanglement might only occur in beds of very small particles, which seems to agree with the

in silico air-free duplications of the effect. But this leaves us with enduring difficulties in understanding the interplay between intruder size and density — and now bed particle size and vibration speed — and

why, to begin with, convection does not prevail to entrain intruders irrespective of such details.

The RBN effect is sure to provide fruit for future exploration and debate. We find ourselves facing the situation anticipated by Mark Twain: “The researches of many commentators have already thrown much darkness on this subject, and it is probable that, if they continue, we shall soon know nothing at all about it.” Although farmers can count on continuing harvests of heavy boulders under any segregation model, it remains to be seen whether — and how — pharmaceutical engineers should expect their granular formulations to mix or separate. ■

Troy Shinbrot is in the Department of Biomedical Engineering, Rutgers University, Piscataway, New Jersey 08854, USA.

e-mail: shinbrot@soemail.rutgers.edu

1. Walker, J. *The Flying Circus of Physics* 72 (Wiley, New York, 1975).
2. Muzzio, F. J., Glasser, B. J. & Shinbrot, T. *Powder Technol.* **124**, 1–7 (2002).
3. Huerta, D. A. & Ruiz-Suárez, J. C. *Phys. Rev. Lett.* **92**, 114301 (2004).
4. Knight, J. B., Jaeger, H. M. & Nagel, S. R. *Phys. Rev. Lett.* **70**, 3728–3731 (1993).
5. Ehrichs, E. E. *et al. Science* **267**, 1632–1634 (1995).
6. Gallas, J. A. C., Herrmann, H. J., Pöschel, T. & Sokolowski, S. *J. Stat. Phys.* **82**, 443–450 (1996).
7. Shinbrot, T. & Muzzio, F. J. *Phys. Rev. Lett.* **81**, 4365–4368 (1998).
8. Breu, A. P. J., Enssner, H.-M., Kruelle, C. A. & Rehberg, I. *Phys. Rev. Lett.* **90**, 014302 (2003).
9. Hong, D. C., Quinn, P. V. & Luding, S. *Phys. Rev. Lett.* **86**, 3423–3426 (2001).
10. Möbius, M. E., Lauderdale, B. E., Nagel, S. R. & Jaeger, H. M. *Nature* **414**, 270 (2001).
11. Canul-Chay, G. A., Belmont, P. A., Nahmad-Molinari, Y. & Ruiz-Suárez, J. C. *Phys. Rev. Lett.* **89**, 189601 (2002).
12. Brone, D. & Muzzio, F. J. *Phys. Rev. E* **56**, 1059–1063 (1997).
13. Yan, X., Shi, Q., Hou, M. & Lu, K. *Phys. Rev. Lett.* **91**, 014302 (2003).

Genome sequencing

Differences with the relatives

Jean Weissenbach

One of the chimpanzee's chromosomes has been sequenced to near-completion. What can this accomplishment tell us about how we have come to look and act so differently from our chimp relatives?

There are good reasons to continue the endeavour to accumulate genome sequence data from the passengers of Noah's Ark. As illustrated on page 382 of this issue¹, genome sequences can serve to address basic evolutionary issues — the power of this approach depending to a large extent on the amount and quality of data available.

The rationale for sequencing the genome of the chimpanzee (*Pan troglodytes*; Fig. 1, overleaf) has been explained on numerous occasions (see ref. 2 for a review), and a publicly funded effort, involving some of the large US sequencing centres, has already produced a draft assembly of the whole

sequence³. But this initial assembly still contains many gaps and ambiguities that present difficulties for some types of analysis.

In an independent effort¹, a consortium of Old World humans has now sequenced chimpanzee chromosome 22 to a degree of completion and accuracy equivalent to that of the human genome assembly in its present version. The quality of this chimp chromosome sequence is therefore good enough to allow reliable comparisons with its human counterpart (chromosome 21). A chimpanzee chromosome provides a unique angle from which to look at the human genome and to draw conclusions about its recent evolution, because the sequences of

these evolutionary near-neighbours started drifting apart some six million years ago. The longer-term hope, of course, is to identify those sequence changes that could account for the present-day physical, physiological and behavioural differences between chimps and people.

By lining up chimp chromosome 22 and human chromosome 21 and comparing them nucleotide by nucleotide, the consortium found instances in which one nucleotide was substituted for another in only about 1.44% of the sequence. The chimpanzee chromosome has been sequenced to an accuracy of less than one error in 10^4 bases, so sequencing mistakes account for less than 1% of the observed single-nucleotide mismatches. There is also an impressive number (68,000) of small to large stretches of DNA that have been either gained or lost (these are called 'insertions or deletions', 'indels' for short) in one species or the other.

The number of single-nucleotide substitutions is in the range found in earlier studies, but the frequency and size of the indels are more of a surprise. Although most of the indels are less than 30 nucleotides long, some attain sizes of up to 54,000 nucleotides. Those of about 300 nucleotides or more frequently involve transposable elements — DNA sequences that multiply and insert new copies of themselves throughout a genome. For a subset of these 300-nucleotide-plus indels, the authors were able to extend the comparison to other great apes: gorillas and orang-utans. They could thereby infer the lineage (chimp or human) in which the alterations occurred, and could distinguish between insertions and deletions — that is, whether a given sequence was added in one lineage or deleted in the other. These comparisons show that insertions of about 300 nucleotides, mainly of the type of transposable element known as an Alu repeat, have occurred preferentially in the human lineage. Deletions and other insertions seem to have occurred at similar frequencies in both lineages.

One of the strongest arguments in support of the chimpanzee genome project was always that having the chimp sequence makes it possible to determine which variants of single nucleotide polymorphisms (SNPs) — single nucleotide differences between individual humans — represent the 'original' form⁴. This information is important in, for instance, genetic association studies aimed at mapping the locations of gene variants associated with complex diseases such as diabetes or high blood pressure. On the basis of the chimp chromosome 22 sequence, the consortium determined the



Figure 1 What makes us different, genetically speaking, from a chimp? The sequencing of chimp chromosome 22 is a step on the way to an answer.

ancestral form of some 20,000 SNPs from human chromosome 21 (although, given that most of the chromosome 22 sequence has come from just one chimpanzee, it remains formally possible that some of the same polymorphisms also occur in chimp populations). The comparison shows that, as expected, transitions (mutation of one purine nucleotide, adenine or guanine, to the other, or of one pyrimidine nucleotide, cytosine or thymine, to the other) are more frequent SNPs than are transversions (mutations of a purine to a pyrimidine and vice versa). Moreover, mutation occurs at guanines and cytosines more frequently than at adenines and thymines.

In searching for the basis of the physical variation between chimps and humans, differences in genome sequences are just the first place to start: we then need to know what these differences mean. Many of the sequence variations might have no effect at all. Those that do might occur in non-protein-coding parts of the genome or in control genes, thereby influencing the level, location and timing of gene expression. Other changes might alter the sequences of encoded proteins, resulting in loss or gain of function. And entire genes might be deleted or acquired in one lineage or another.

Given the broad similarities between chimps and humans, many researchers thought that changes that alter amino-acid sequences would not be very frequent.

Surprisingly, however, the consortium found that sequence differences in the protein-coding regions of genes are not a great deal less common than in non-coding genomic regions. But some of the affected genes might be pseudogenes — defective copies of functional genes — that have arisen recently. And, among 231 presumably functional genes that could be compared between chimps and humans, 179 have protein-coding regions of identical length; 140 of the predicted encoded proteins would differ by one amino acid or more, but probably with little or no functional impact. Of the other 52 genes, however, 47 show more significant structural changes.

The consortium could not resist making preliminary studies of the expression of the genes on human chromosome 21 and chimp chromosome 22 as well. Their analyses indicate that — looking at just two tissues — about 20% of these genes show significant variations in their expression. Extrapolation from these findings suggests that if this chromosome represents about 1% of mammalian genes, there may well be thousands of genes that either encode an altered protein or are expressed differentially in humans and chimpanzees. This

will not simplify the search for the hypothetical key genetic changes that prevented us from remaining as apes.

Even if the major physical, physiological and behavioural differences between the two species do not result simply from an accumulation of many small alterations, the challenge to find the most crucial changes is still ahead. For example, the FOXP2 gene product, which is important for language development, differs by two amino acids in humans and chimps, suggesting that the gene has been a target of selection in the human lineage. Yet the role of this gene in language was suggested not by human–chimp comparisons⁵ but by mutation studies in humans⁶.

Identifying sequence changes in the chimpanzee lineage that are likely to have been irrelevant to the acquisition of human-specific traits will depend on sequence comparisons with other great apes. Do we now need the gorilla genome sequence to shed more light on the questions raised by comparing human and chimp DNA? ■

Jean Weissenbach is at Genoscope, 91000 Evry, France.

e-mail: jsbach@genoscope.cns.fr

1. The International Chimpanzee Chromosome 22 Consortium *Nature* **429**, 382–388 (2004).
2. Olson, M. & Varki, A. *Nature Rev. Genet.* **4**, 20–28 (2003).
3. www.genome.gov/11509418
4. Hacia, J. G. et al. *Nature Genet.* **22**, 168–171 (1999).
5. Enard, W. et al. *Nature* **418**, 869–872 (2002).
6. Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. *Nature* **413**, 519–523 (2001).

Earth science

Just add more water

Rob Evans

Data from a different technique for probing events within the Earth, interpreted in terms of a new hypothesis about the effects of water at depth, raise tantalizing questions about recycling of a tectonic plate.

In places such as the Pacific Rim, the sea floor collides with and is pushed below the land masses that border the ocean. This process of subduction, and its consequences in the form of earthquakes and volcanism, are the result of plate tectonics, in which the formation of sea floor at mid-ocean ridges is accommodated by the recycling of tectonic plates in the Earth's interior. But the recycling is a complex process: the interaction of the subducting plate, or slab, with the surrounding material itself produces volcanism and, in some settings, new sea floor. On page 399 of this issue, Booker *et al.*¹ present dramatic models of the subduction system beneath the Andes that suggest that the recycling processes run deeper than previously thought. Their interpretation depends on new data, which in themselves are striking but which also allow the authors to make some inferences about the influence of water at depth.

Water plays an important role in subduction recycling: its release from the oceanic plate can affect the rheology of the rock as it moves into the mantle, below the Earth's crust, and also increases rock melting². Yet the amount of water released, the depth of release and the pathways it takes are poorly defined, as are the depths of melting and the route that the melt takes as it moves from close to the subducting plate to its eruption in the overlying volcanic arc.

Seismic techniques for looking into the Earth will be familiar to most people. But there are other approaches, one being the magnetotelluric (MT) method, which offers promise for locating melt and identifying the distribution of water in the mantle. This technique uses naturally occurring electric currents in the ionosphere, created by the capture of charged particles by the planet's magnetic field, to measure Earth's electrical conductivity. That conductivity depends partly on composition and temperature. But it can be dramatically increased by small amounts of partial melt, provided that the melt forms an interconnected network³. And it can also be affected by water in the mantle, in the form of dissolved hydrogen, both above and below the transition that occurs globally at around 410 km, where the mantle composition undergoes a change from one mineral phase (olivine⁴) to another (wadsleyite⁵). MT has been used only infrequently to address subduction-zone processes, mostly because these systems are close to the ocean and require complex onshore-offshore investigations^{6,7}. Not only are the results of Booker *et al.*¹ dramatic in their own right, they also represent a significant advance in MT imaging of subduction systems.

Where Booker *et al.* have made their measurements, the subducting slab flattens beneath South America before

plunging down into the mantle. This makes the area somewhat anomalous — an inactive volcanic arc sits about 200 km above the slab, instead of the more normal 100 km. But it does mean that the key subduction processes occur farther inland beneath the continent, and can be imaged using terrestrial MT (see Fig. 1 of the paper on page 400).

The most significant part of Booker and colleagues' model is the conductive anomaly that appears to coincide with the top of the subducting slab. The introduction of fluids from the slab, resulting in melting, seems one obvious explanation for the anomaly. Yet what is surprising is the great depth — at least 250 km but perhaps as much as 400 km — to which this melting seems to be occurring. Earthquake activity from the adjacent slab (there is very little activity in the region sampled by Booker *et al.*) shows intermediate-depth earthquakes ending around 200 km, and it has been suggested that this marks the depth limit of water-bearing mineral phases^{8,9}. In fact, the MT data indicate that there is electrical conductivity along or just above the flat part of the slab (above 200 km), suggesting some water loss at these depths.

There are few other clues that can help to account for this apparent deeper melting. Analyses of samples from an overlying inactive volcano, Pocho, are equivocal, because much of the chemistry of the samples reflects processes occurring in the crust¹⁰; in any event, their age means that they might easily reflect a different melting geometry. The prolonged passage of an old slab under the continent at relatively shallow depths is unusual, and might provide an explanation for why the slab has been able to hold on to its water to greater depths than usual. Perhaps it results in the breakdown of mineral phases that do not normally contribute to the water budget.

Organic chemistry

000!

In an organic molecule, how long can a chain of oxygen atoms be? No chain-like polyoxides are known that have more than three oxygen atoms in a row; only a few trioxides are known. But to this collection, Holger Pernice *et al.* now add bis(fluoroformyl)trioxide, or FC(O)OOC(O)F (*Angew. Chem. Int. Edn* **43**, 2843–2846; 2004).

Bis(fluoroformyl)trioxide is an acyl compound, characterized by a C=O acid-group derivative. It is the second known example of a chain-like acyl trioxide (after the more complex $\text{CF}_3\text{OC(O)OOC(O)OCF}_3$) — but the first to have its twisting, chiral structure determined.

Starting from a potentially explosive mixture of carbon monoxide, fluorine and oxygen, the reaction proceeds to produce the acyl trioxide and the peroxide FC(O)OOC(O)F . The spectrum of infrared radiation absorbed by the sample has bands at around 800 and 900 cm^{-1} that are typical of O-O-O stretching vibrations. Similarly, data from ultraviolet, nuclear magnetic resonance and X-ray diffraction studies confirm the formation of the trioxide.

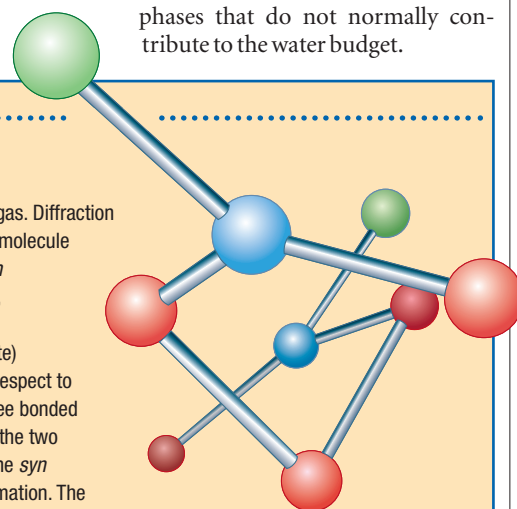
For the determination of its structure, Pernice *et al.* allowed a trioxide sample to crystallize at a temperature of -78°C in a

stream of nitrogen gas. Diffraction data show that the molecule has a *trans-syn-syn* structure — that is, the C-O bonds are in the *trans* (opposite) conformation with respect to the plane of the three bonded oxygen atoms, and the two C=O bonds are in the *syn* (same-side) conformation. The left-handed version of this structure is shown here (oxygen in red, carbon in blue, fluorine in green).

Theory predicts that another structure, or 'rotamer', should exist: *trans-syn-anti*, in which, instead of curling round, the molecule

twists back on itself about the central oxygen atom. Pernice *et al.* say that their infrared data indeed suggest the presence of a second rotamer.

Alison Wright



Booker *et al.*¹ explain the deep melting by invoking the ideas of Bercovici and Karato¹¹. This new hypothesis suggests that a layer of melt exists just above the 410-km boundary as slowly upwelling mantle undergoes a change from water-rich wadsleyite to olivine, liberating water and inducing melting. In Booker and colleagues' models, the melt column seems to originate as much from the 410-km boundary as it does from the subducting slab, and so the suggestion of a link between the two is not surprising. The authors propose that the liberation of additional water from the slab induces further production of buoyant melt, a necessary condition if the melt is to rise as is seen.

This link to Bercovici and Karato's model is intriguing, but I'm not completely sold on it. Undoubtedly, the argument would be strengthened by better data on the structure across the 410-km boundary which, as the authors point out, is not the best-resolved feature of the model. Booker *et al.* suggest that the mantle to the west of (or beneath) the slab is dehydrated as a result of melt extraction at the East Pacific Rise, part of the mid-ocean ridge system in the Pacific Ocean. Although this might be true for the upper 60–80 km or so of oceanic plate from which melt and water have been extracted, other data suggest that there is plenty of remaining water that can increase electrical conductivity in the mantle below about 80 km depth¹². This means that only the region adjacent to the subducting slab would be expected to be dry. If there is water around, we would expect a more or less uniform increase in conductivity at 410 km across the region. Although their models show a stepwise increase across the 410-km boundary, Booker *et al.* point out that the data are consistent with a flat 410-km transition throughout the region. Why is all this important? Well, it speaks to outstanding issues in terms of resolution of this critical part of the system.

Regardless of the details of the 410-km boundary, the authors' primary observation — an electrical conductor that must surely represent a subduction-related melt column rising from depth — is striking. And, as they point out, issues pertaining to the 410-km boundary and the link to the Bercovici–Karato hypothesis can be addressed with measurements made with a longer chain of MT stations. If the interaction with a melt layer at 410 km is the explanation, it should be seen in other subduction systems. It hasn't been seen elsewhere yet, but maybe we just need to look more carefully.

Rob Evans is at Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. e-mail: revans@whoi.edu

1. Booker, J. R., Favetto, A. & Pomposiello, M. C. *Nature* **429**, 399–403 (2004).
2. Hirth, G. & Kohlstedt, D. L. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
3. Roberts, J. J. & Tyburczy, J. A. *J. Geophys. Res.* **104**, 7055–7066 (1999).

4. Karato, S. *Nature* **347**, 272–273 (1990).
5. Xu, Y. *et al. Science* **280**, 1415–1418 (1998).
6. Wannamaker, P. E. *et al. J. Geophys. Res.* **94**, 14127–14144 (1989).
7. Echternacht, F. *et al. Phys. Earth Planet. Inter.* **102**, 69–87 (1997).
8. Kirby, S. *et al. AGU Geophys. Monogr.* **96**, 195–214 (1996).
9. Hacker, B. R. *et al. J. Geophys. Res.* **108**, doi:10.1029/2001JB001129 (2003).
10. Kay, S. M. & Gordillo, C. E. *Contrib. Mineral. Petrol.* **117**, 25–44 (1994).
11. Bercovici, D. & Karato, S. *Nature* **425**, 39–44 (2003).
12. Lizarralde, D. *et al. J. Geophys. Res.* **100**, 17837–17854 (1995).

Ageing

Mice and mitochondria

George M. Martin and Lawrence A. Loeb

It can be hard to work out whether particular events are a cause or a correlate of ageing — do mutations in mitochondrial DNA, for instance, speed up the process of growing old? Some clever studies suggest so.

Mitochondria are little pockets of energy within cells, shouldering the important task of converting food into usable energy forms. To meet demands, every cell contains thousands of them. Unlike most other cellular compartments, mitochondria have their own genomes, which encode a few mitochondrial proteins (most others being encoded by genes in the nucleus). In numerous non-reproductive tissues of many species, mitochondrial genes (like nuclear genes) accumulate mutations as the animals age¹, and it has been speculated that these mutations might in fact cause ageing, by leading to energy-generation defects — increased numbers of harmful reactive oxygen species, cellular damage and so on. On the other hand, the association between mitochondrial mutations and ageing could merely be correlative — these mutations might simply be one of the manifestations of growing old. That possibility becomes less likely, however, with the publication of the paper on page 417 of this issue by Trifunovic and colleagues².

Point mutations (single base-pair changes)³, deletions and rearrangements⁴ in mitochondrial DNA accumulate in several non-reproductive (somatic) tissues during ageing. To find out whether such alterations are a cause or a correlate of growing old, Trifunovic *et al.*² genetically engineered mice to carry mutations in an enzyme called DNA polymerase- γ . Encoded by nuclear genes and transported to mitochondria, this protein is

believed to be responsible for all aspects of mitochondrial DNA metabolism: it both copies and proofreads the DNA, eliminating errors it makes during replication, and it is also believed to participate in the resynthesis of DNA during DNA-repair processes. New mitochondria replace old mitochondria in all cell types throughout life, and new mitochondria must also be made when cells divide. These events require the replication of mitochondrial DNA.

Trifunovic and colleagues wanted to render the mitochondrial DNA polymerase error-prone by eliminating its proofreading activity while maintaining its catalytic potency — the rationale being that any errors in mitochondrial DNA replication would go unnoticed by the cell, and so mutations would accumulate. As the mice would have this error-prone polymerase from youth, it would be possible to see whether or not the mutations it produces accelerate ageing. To eliminate the proofreading activity, the authors substituted an alanine amino acid for a crucial aspartate in the relevant region of the enzyme molecule.

They found that the somatic tissues of mice bearing two mutant copies of the DNA polymerase- γ gene indeed showed extensive mitochondrial DNA mutations, largely comprising deletions and point mutations. The percentage of mitochondria bearing deletions was similar in different tissues, and did not vary with age, suggesting that the deletions had occurred early in development.

Mutated protein	Effects
Lamin A/C (ref. 11)	Defective inner nuclear membrane
Ku86, XPD (refs 12, 13)	Defective metabolism of nuclear DNA
DNA polymerase- γ (ref. 2)	Defective metabolism of mitochondrial DNA
Telomerase (ref. 14)	Defective regulation of chromosome caps
p53 (ref. 15)	Altered regulation of cell-division cycle and cell death
Klotho (ref. 16)	Impaired calcium and vitamin D metabolism?

Figure 1 Many mutational pathways can accelerate 'segmental ageing' (or, to use more conservative nomenclature, segmental ageing-like syndromes¹⁰) in mice. Published literature^{2,11–16} suggests that there may be at least six pathways that generate partially overlapping subsets of features consistent with accelerated ageing. Several of these pathways are likely to interact¹⁷, making such classifications problematic. Trifunovic and colleagues' findings², however, provide strong direct support for the mitochondrial pathway.

The deletions seemed to have converted the normally circular mitochondrial DNA into linear molecules, which might not be functional inside cells.

Point mutations were also common; one mitochondrial enzyme, for instance — cytochrome *b* — had a three- to fivefold increase in single-base substitutions, randomly dispersed throughout the gene. The mutant animals also showed a decrease in the activity of enzymes involved in the respiratory chain (a crucial series of events in respiration) and the production of ATP (the main cellular energy store), which could be a result of both the deletions and the point mutations. The mutant mitochondrial DNA molecules were present together with normal copies, but sometimes a particular mutation would come to dominate — a situation comparable to that seen in ageing humans⁵.

Strikingly, the mutant mice showed symptoms consistent with accelerated ageing, such as premature weight loss, hair loss, reductions in fertility, curvature of the spine and a shortened lifespan. The symptoms began to emerge at about 25 weeks of age — young adulthood, in mouse terms. These findings strongly support the idea that mutations in mitochondrial DNA can cause at least some features resembling ageing. This delayed, post-maturational expression of overt signs of ageing is an important feature of any useful model system for studying the mechanisms by which organisms grow old.

Trifunovic and colleagues' findings are also consistent with the oxidative-damage theory of ageing — the idea that ageing is caused by an increase in the steady-state levels of reactive oxygen species and by proton 'leaks'. Such events can result from mutations in respiratory-chain proteins.

Interestingly, there is a human disease that might be considered as a parallel to these mutant mice. Like mouse DNA polymerase- γ , the human enzyme contains DNA-synthesizing and proofreading domains⁶. Mutations in the proofreading domain are among the genetic alterations responsible for a rare inherited human disease, progressive external ophthalmoplegia⁷, which is characterized by paralysis of the eye muscles and various effects on other organs, and by the accumulation of mitochondrial mutations in non-reproductive tissues⁸. Like other mitochondrially linked, inherited disorders, the disease typically exhibits a delayed onset and a progressive course — features shared with ageing.

As with the mutant mice, however, many characteristics of interest to gerontologists have not been investigated in the rare affected people. More information is needed, in both mice and humans, about the ages of onset and rates of progression of cataracts, visual degeneration and hearing loss, and changes in immunity, hormones, cognitive functions and other traits.

Nonetheless, Trifunovic and colleagues' elegant study² establishes that several manifestations of ageing in mice can result from mutations in DNA polymerase- γ that induce error-prone mitochondrial DNA synthesis. These results do not, however, imply that this is the only pathway that generates abnormalities resembling ageing. There are several other DNA-replicating and proofreading polymerases (α , β and δ) that function in the nucleus, as well as at least eight newly discovered, naturally error-prone polymerases that are believed to function in bypassing nuclear DNA lesions or in specialized processes that affect DNA⁹. These nuclear enzymes might, when mutated, each lead to further genetic alterations in certain somatic tissues and so accelerate ageing. If experiments show this to be the case, these polymerases will join the growing list of proteins that, when mutated in mice, produce groups of features that hint at an acceleration of particular aspects of ageing ('segmental ageing'; Fig. 1); many of the mutations could act by enhancing genomic instability.

Of higher priority, however, would be experiments aimed at finding ways of maintaining the structure and function of tissues and organs for longer periods — leading to

lengthened, not abbreviated, lifespans. We therefore look forward to the availability of mice that have been modified to bear a mitochondrial DNA polymerase that is more accurate than the normal enzyme. ■

George M. Martin and Lawrence A. Loeb are in the Department of Pathology, University of Washington, Seattle, Washington 98195, USA.

e-mails: gmartin@u.washington.edu

laloeb@u.washington.edu

1. Vijg, J. *Mutat. Res.* **447**, 117–135 (2000).
2. Trifunovic, A. *et al. Nature* **429**, 417–423 (2004).
3. Wang, Y. *et al. Proc. Natl Acad. Sci. USA* **98**, 4022–4027 (2001).
4. Melov, S., Hinerfeld, D., Esposito, L. & Wallace, D. C. *Nucleic Acids Res.* **25**, 974–982 (1997).
5. Nekhaeva, E. *et al. Proc. Natl Acad. Sci. USA* **99**, 5521–5526 (2002).
6. Lim, S. E., Longley, M. J. & Copeland, W. C. *J. Biol. Chem.* **274**, 38197–38203 (1999).
7. Agostino, A. *et al. Neurology* **60**, 1354–1356 (2003).
8. Del Bo, R. *et al. Neurology* **61**, 903–908 (2003).
9. Shcherbakova, P. V., Bebenek, K. & Kunkel, T. A. *Science* **301**, 122–124 (2003). doi:10.1126/science.1083833
10. Martin, G. M. *Birth Defects Orig. Artic. Ser.* **14**, 5–39 (1978).
11. Mounkes, L. C., Kozlov, S., Hernandez, L., Sullivan, T. & Stewart, C. L. *Nature* **423**, 298–301 (2003).
12. Vogel, H., Lim, D. S., Karsenty, G., Finegold, M. & Hasty, P. *Proc. Natl Acad. Sci. USA* **96**, 10770–10775 (1999).
13. De Boer, J. *et al. Science* **296**, 1276–1279 (2002).
14. Rudolph, K. L. *et al. Cell* **96**, 701–712 (1999).
15. Tyner, S. D. *et al. Nature* **415**, 45–53 (2002).
16. Nabeshima, Y. *Ageing Res. Rev.* **1**, 627–638 (2002).
17. Orsini, F. *et al. J. Biol. Chem.* doi:10.1074/jbc.M401844200 (2004).

Geochemistry

Warm debate on early climate

Timothy W. Lyons

Would Earth's early ocean have been a frozen wasteland had levels of atmospheric methane not been sky high? Maybe. Or maybe, according to a new view of an old idea, the main warming agent was carbon dioxide.

For the first three-and-a-half billion years of Earth's history, the Sun burned only about 70–90% as brightly as it does today. A famous paradox of ancient climate demands that we reconcile the persistence of a life-sustaining liquid ocean with these less-warming rays from a faint young Sun. The solution lies with greenhouse gases¹, which trap heat near the planet's surface. In recent years, methane has become the favoured greenhouse agent, overcoming the earlier popularity of carbon dioxide in climate models spanning most of the Precambrian eon — that is, during all of Earth's history up to about three-quarters of a billion years ago, after which life emerged on a large scale.

On page 395 of this issue, Ohmoto and his colleagues² dispute the arguments against CO₂. Their challenge has additional significance because the assertion of inadequate CO₂ is the most frequently cited evidence for the existence of high levels of atmospheric methane. And given the incompatibility of methane and oxygen, this debate speaks more broadly to the oxygenation history of

the atmosphere and its link to the evolution of early life.

Today, most of us worry about the 30% rise in levels of CO₂ since the Industrial Revolution and how that may drive global warming. But a generation of models of Earth's atmosphere invoked CO₂ concentrations as high as a thousand times greater than those of today to explain the unfrozen early ocean¹. About a decade ago the idea of CO₂ as the dominant greenhouse agent was dealt a blow when Rye and his co-workers³ used the absence of siderite, an iron-rich carbonate mineral, in ancient soils to set a maximum for CO₂ levels in the atmosphere 2.2 billion to 2.75 billion years ago (Fig. 1, overleaf). This maximum, although still many times the concentration observed today, was well below that necessary to offset the faint young Sun.

Rye *et al.* were compelled to suggest that another greenhouse gas — methane — must have taken up the slack. But their story was based on only a handful of data and assumed, among other things, that original mineral constituents and chemical properties

can be inferred from the rocky remnants of highly altered ancient soils. Also, chemical reactions in the natural world frequently deviate from the equilibrium behaviour assumed in Rye and colleagues' thermodynamic model.

Nonetheless, the absence of siderite in ancient soils has become the foundation of arguments for lower CO₂ and thus for a methane-rich early atmosphere. Precious little direct evidence backs up that view, which is based mainly on the CO₂ estimates, but a body of independent, circumstantial evidence lends compelling support. By most accounts, for example, the atmosphere was deficient in oxygen before about 2.3 billion years ago⁴. Anoxia favours methane preservation, as well as methane production by methanogens — anaerobic, or oxygen-intolerant, microorganisms that first appeared early in the Precambrian.

Supplies of sulphate in the ocean also figure prominently in the production and longevity of methane. Sulphate is the second most abundant anion in sea water today. But because it is produced mostly by weathering on the continents in the presence of oxygen, the early ocean would have been sulphate poor⁵. Bacteria that reduce sulphate compete with methanogens for metabolizable organic compounds, and anaerobic decomposition of methane requires the presence of sulphate⁶. Conditions that were comparatively

favourable to methane, including low concentrations of sulphate in sea water, may have persisted until about 0.75 billion years ago^{7,8}, when oxygen concentrations in the atmosphere and ocean once again took a big step upward.

Ohmoto *et al.*² have a different take on the presence or absence of iron-rich carbonate minerals in ancient soil horizons. By taking a broader thermodynamic view of these so-called palaeosols, they suggest that siderite is absent largely because oxygen levels were too high. Ohmoto *et al.* are quick to point out that even minuscule levels of oxygen could have been enough to disfavour siderite formation. As with the model of Rye *et al.*, this conclusion requires assumptions about equilibrium chemical conditions, ambient temperatures, and gas partitioning between the atmosphere and subsurface soil environments — among other poorly known parameters. If the assumptions of Ohmoto *et al.* are correct, this additional constraint tells us that CO₂ levels in the atmosphere could have been higher than those suggested by Rye and his colleagues.

Despite the absence of siderite in palaeosols, which Ohmoto *et al.* attribute to factors other than CO₂, the mineral was common in marine settings before 1.8 billion years ago. Under these conditions, Ohmoto *et al.* argue, the controlling factor was CO₂, implying that levels of this gas in the atmosphere must have been high — high enough to minimize the need to invoke methane as a major greenhouse gas.

As with the palaeosols, however, factors other than CO₂ availability can control the formation of siderite in marine environments. For example, the high concentrations of dissolved iron that most workers imagine

existed in the early, oxygen-deficient ocean⁹ would reduce the amounts of CO₂ required for the mineral to form. Given such uncertainties, it is difficult to view Ohmoto and colleagues' estimates of high CO₂ levels as the only interpretation of early distributions of siderite. Furthermore, although concentrations of CO₂ higher than those proposed by Rye *et al.* may weaken the principal argument for high methane concentrations, they don't preclude both CO₂ and methane from being much more plentiful than they are in today's atmosphere^{10,11}.

A universal theme in studies of the early Earth is that big stories are told with little data and lots of speculation, and interpretations of Precambrian siderite are no exception. Nevertheless, Ohmoto *et al.* open our eyes to another way of looking at the presence and absence of this diagnostic mineral. In doing so, they show us that exploratory steps can lead in many directions, but with each step the ancient atmosphere becomes a little more transparent.

Timothy W. Lyons is in the Department of Geological Sciences, University of Missouri, Columbia, Missouri 65211, USA.

e-mail: lyons@missouri.edu

1. Kasting, J. F. *et al. Science* **259**, 920–926 (1993).
2. Ohmoto, H., Watanabe, Y. & Kumazawa, K. *Nature* **429**, 395–399 (2004).
3. Rye, R., Kuo, P. H. & Holland, H. D. *Nature* **378**, 603–605 (1995).
4. Holland, H. D. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).
5. Habicht, K. S., Gade, M., Thamdrup, B., Berg, P. & Canfield, D. E. *Science* **298**, 2372–2374 (2002).
6. Boetius, A. *et al. Nature* **407**, 623–626 (2000).
7. Pavlov, A. A., Hurtgen, M. T., Kasting, J. F. & Arthur, M. A. *Geology* **31**, 87–90 (2003).
8. Kah, L. C., Lyons, T. W. & Chesley, J. T. *Precamb. Res.* **111**, 203–234 (2001).
9. Canfield, D. E. *Nature* **396**, 450–453 (1998).
10. Pavlov, A. A., Kasting, J. F., Brown, L. L., Rages, K. A. & Freedman, R. J. *Geophys. Res.* **105**, 11981–11990 (2000).
11. Hessler, A. M., Lowe, D. R., Jones, R. L. & Bird, D. K. *Nature* **428**, 736–738 (2004).



Figure 1 The answers lie in the soil? Right, the Pilbara region of northwestern Australia, which is famous for its old rocks and their geochemical signatures of conditions on the surface of the early Earth. The Mount Roe palaeosols, ancient soil horizons in this region dating from about 2.75 billion years ago (above), have featured prominently in the debate over which greenhouse gas — methane or carbon dioxide — predominated in keeping the ancient Earth warm.



Interplanetary ecology

Niche prospects on Mars

Icarus **169**, 300–310 (2004)

Microorganisms that harness solar energy must gather enough visible light to carry out photosynthesis while avoiding the DNA-damaging effects of ultraviolet (UV) radiation. Today, Earth is shielded from UV radiation by the ozone layer. But these microorganisms probably arose without such protection — raising the intriguing question of whether there are places on harsh worlds, such as Mars, where microbes might get enough light while being sheltered from UV radiation.

Charles S. Cockell and John A. Raven propose that there may indeed be such niches on Mars. Their conclusion is based on a model of how radiation travels through the martian atmosphere, along with lab observations of the way light penetrates various substrates.

Compared with Earth, the martian surface receives UV radiation up to a thousand times more damaging to DNA, and only 55% of the visible light. Nonetheless, photosynthetic organisms could potentially flourish under a few millimetres of iron-containing sediment, rock or salts, or a few centimetres of snow, the authors calculate. They brand these regions 'martian Earth-like photosynthetic zones', and suggest that they resemble the regions where Earth's sun-loving microbes arose billions of years ago.

Michael Hopkin

Developmental biology

End extension

Proc. Natl Acad. Sci. USA
doi:10.1073/pnas.0402400101 (2004)

Telomeres — the ends of chromosomes — become frayed during life and are restored when sperm and eggs are made. But there has been conflicting evidence about whether they are also repaired in cloned animals, which bypass normal fertilization. Sonja Schaezlein *et al.* show that, in some cases at least, telomeres are indeed restored during early embryonic development following cloning.

The authors measured the length of telomeres in cow fibroblast cells whose nuclei were used for cloning, and in embryos produced by cloning or conventional breeding. They used a fluorescent probe that attaches to telomere-specific genetic sequences: the longer the telomere, the brighter the fluorescent signal. Following cloning, telomeres were shorter than normal when the embryos were still a compact ball of cells called a morula. But later, when the embryos had expanded into a blastocyst, their telomeres were restored to normal

length. Regular cow and mouse embryos also showed telomere lengthening during the transition between these stages.

Because decaying telomeres are thought to contribute to cell ageing, Schaezlein *et al.* suggest that deciphering and reactivating this elongation programme might help to prolong the life of cells used in treatments. It remains unclear why telomere repair does not occur in cloned animals derived from some other cell types.

Helen Pearson

Analytical chemistry

A faster pollution detector

Anal. Chem. doi:10.1021/ac0498260 (2004)

A specially modified mass spectrometer can detect urban pollutants with much greater sensitivity and speed than rival methods. Robert S. Blake and colleagues report that their combination of a proton-transfer ionization reaction with time-of-flight mass spectrometry has improved detection sensitivity more than 100-fold, such that trace volatiles can be picked up at concentrations of parts per billion. Unlike other systems, complex mixtures can be analysed in under a minute, and the mass range is not limited to smaller molecules.

The analysis begins with the ionization of water vapour, to create H_3O^+ ions, by α -particles emitted from a radioactive americium source. H_3O^+ ions are the proton carriers found in aqueous acids, and will transfer H^+ to any molecule that has a higher proton affinity than water. This includes most of the volatile organic compounds that are important trace pollutants, but excludes the major components of air such as oxygen and carbon dioxide, thus avoiding problems of strong background signals. The protonated organic molecules, now positively charged, move through an electric field until they hit a detector. Their time of flight is measured to calculate their mass.

The authors claim that this technique could be used to detect the small quantities of numerous heavy, volatile organic compounds that may make a substantial, but as yet unquantified, contribution to airborne pollution.

Mark Peplow

Cancer

Fishing in the ribosome

PLoS Biol. **2**, 0690–0698 (2004)

Zebrafish are emerging as a promising model organism in cancer research. The latest news comes from Adam Amsterdam *et al.*, who show that in these fish many protein components of the ribosome, the site of protein synthesis, apparently function as tumour suppressors — that is, cancer is more likely to develop if they are absent.

Amsterdam *et al.* screened various lines of mutant zebrafish and found that, in



several of them, mutations in genes encoding ribosomal proteins resulted in the development of tumours and subsequent early death. A systematic survey of other lines carrying mutations in different ribosomal proteins revealed that a high percentage were cancer prone, mostly developing malignant tumours of the sheaths that surround peripheral nerves.

It remains to be seen how mutations in ribosomal proteins predispose zebrafish to cancer. But the next step will be to see if a similar form of tumour suppression occurs in humans — as the authors point out, the action of mutated ribosomal-protein genes as 'cancer genes' may well have been overlooked.

Barbara Marte

Biochemistry

An a-peeling idea

J. Agric. Food Chem. **52**, 2879–2886 (2004)

The peel of citrus fruits contains health-promoting chemicals called polymethoxylated flavones (PMFs), which are known for their anti-cancer and anti-inflammatory effects. A study in hamsters by Elzbieta M. Kurowska and John A. Manthey suggests that PMFs from tangerines might also help to lower cholesterol levels.

Kurowska and Manthey fed hamsters on a diet that produces high blood cholesterol levels. Animals that were also given 1% PMFs in their feed experienced a decrease of up to 40% in their concentrations of low-density lipoprotein — 'bad' cholesterol. Levels of 'good' cholesterol (high-density lipoprotein) remained unaltered, and there were no adverse side effects.

Similar results were obtained using a more concentrated supplement of flavanones, another type of chemical found in citrus fruit. So the authors speculate that PMFs have a more potent cholesterol-lowering effect than the flavanone alternative. PMF metabolites were found in the hamsters' livers, hinting that perhaps the chemicals owe their efficacy to their extensive absorption and metabolism. Their therapeutic potential in humans has yet to be tested.

Helen R. Pilcher

The red sweat of the hippopotamus

The red and orange pigments in this secretion account for its protective properties.

P. JOHNSON/CORBIS

Within a few minutes of perspiration, the colourless, viscous sweat of the hippopotamus gradually turns red, and then brown as the pigment polymerizes. Here we isolate and characterize the pigments responsible for this colour reaction. The unstable red and orange pigments turn out to be non-benzenoid aromatic compounds that are unexpectedly acidic and have antibiotic as well as sunscreen activity.

Although the fluid secreted by the hippopotamus (*Hippopotamus amphibius*) is not strictly sweat as it is produced by the subdermal glands¹, it acts like sweat in helping to control body temperature. It is also thought to be antiseptic¹. We collected this highly alkaline (pH 8.5–10.5; ref. 1) red secretion by wiping a hippopotamus's face and back with gauze and extracting its chemical constituents with water. (For details of methods, see supplementary information.)

This diluted red solution was carefully concentrated to about 10^{-5} M; no further concentration was possible as the solution would have polymerized and turned brown. Subsequent purification by gel filtration and on an ion-exchange resin yielded solutions of two pigments, one red and one orange. We characterized these samples by spectroscopy (ultraviolet spectrum, ¹H NMR, electrospray ionization and fast-atom bombardment mass spectroscopy). We also converted the labile red pigment into a stable *t*-butyldimethylsilyl (TBS) derivative and determined its structure by X-ray crystallographic



Cool customer: hippos create their own antibiotic sunscreen.

analysis (Fig. 1, compound 1).

From these results, the structures of the red and orange pigments were deduced (Fig. 1, compounds 2 and 3, respectively). We named the red pigment 'hipposudoric acid' and the orange pigment 'norhipposudoric acid'. The framework of 2 and 3 is probably biochemically derived by oxidative dimerization of homogentisic acid, a metabolite of the amino acids phenylalanine and tyrosine.

The tautomeric or resonant structures of the highly hydrophilic 2 and 3 were studied by using a model compound 4, which shares a chromophore with these pigments and is soluble in organic solvents. Compound 4 was prepared by oxidation of the corresponding tetrahydrodiquinone in deuterated chloroform.

Analysis of the ¹H NMR spectrum of 4

in deuterated chloroform indicates that the structure is symmetrical and that the hydroxy proton of the enol group forms hydrogen bonds with the oxygen atoms at positions 4 and 5 (Fig. 1, 4a) to give a mesomeric diketonate; alternatively, 4 may exist as a mixture of two tautomers.

The pK_a value of 4 in 50% aqueous methanol was measured as about 2.7–3.3; under the same conditions, the pK_a values of acetic acid and phenol are 5.6 and 11.8, respectively. This indicates that 4 is an unexpectedly strong acid that should exist predominantly in its dissociated form (5 in Fig. 1) in water.

On the basis of their structural similarity to 4, the natural pigments 2 and 3 are also likely to exist in their anionic forms in aqueous solution and in the hippopotamus's alkaline sweat, their charge delocalized by resonance. The anions are probably stabilized by intermolecular hydrogen bonding with the solvent.

What is the function of these pigments as far as the hippopotamus is concerned? Their spectra in the ultraviolet/visible range (200–600 nm; see supplementary information) indicate that they may act as sunscreens. The red pigment 2 also has antibiotic activity: at concentrations lower than that found on the hippopotamus's skin, it inhibits the growth of the pathogenic bacteria *Pseudomonas aeruginosa* A3 and *Klebsiella pneumoniae* (see supplementary information).

The isolated pigments are highly unstable, but when they dry on the skin surface in the presence of mucus they maintain their colour for several hours before they polymerize into brown solids. The pigment precursors remain to be identified and the stabilizing effect of the mucus has yet to be investigated.

Yoko Saikawa*, Kimiko Hashimoto*†, Masaya Nakata*, Masato Yoshihara‡, Kiyoshi Nagai‡, Motoyasu Ida‡, Teruyuki Komiyama‡

*Department of Applied Chemistry, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
†Present address: Kyoto Pharmaceutical University, 1 Shichono-cho, Misasagi, Yamashina-ku, Kyoto 607-8412, Japan

e-mail: kimikoh@mb.kyoto-phu.ac.jp

‡Ueno Zoological Gardens, 9-83 Ueno Kouen, Taitou-ku, Tokyo 110-8711, Japan

1. Eltringham, S. K. *The Hippos* 8–38 (Poyser Natural History Series, London, 1999).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

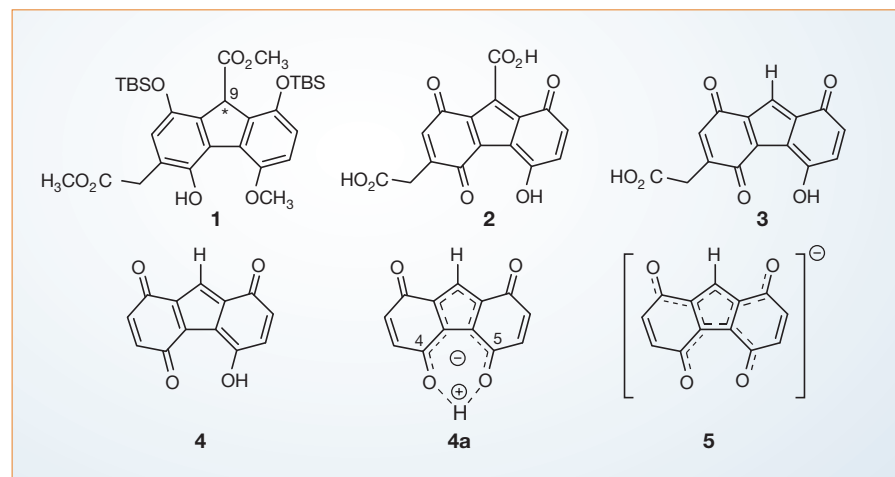


Figure 1 Structures of the red and orange pigments isolated from hippopotamus sweat. Compound 1 is the *t*-butyldimethylsilyl (TBS) derivative of the labile red pigment hipposudoric acid (compound 2) used for X-ray crystallography; asterisk indicates a racemic carbon; compound 3 is the orange pigment norhipposudoric acid. Compound 4 is the synthesized model pigment; 4a is the coordinated hydrogen-bonded form; and 5 is the anionic structure of the model, which resembles the forms of the pigments that exist above about pH3.

Quality assessment of the human genome sequence

Jeremy Schmutz, Jeremy Wheeler, Jane Grimwood, Mark Dickson, Joan Yang, Chenier Caoile, Eva Bajorek, Stacey Black, Yee Man Chan, Mirian Denys, Julio Escobar, Dave Flowers, Dea Fotopulos, Carmen Garcia, Maria Gomez, Eidelyn Gonzales, Lauren Haydu, Frederick Lopez, Lucia Ramirez, James Retterer, Alex Rodriguez, Stephanie Rogers, Angelica Salazar, Ming Tsai & Richard M. Myers

Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, 975 California Avenue, Palo Alto, California 94304, USA

As the final sequencing of the human genome has now been completed, we present the results of the largest examination of the quality of the finished DNA sequence. The completed study covers the major contributing sequencing centres and is based on a rigorous combination of laboratory experiments and computational analysis.

From the beginning, a primary objective of the Human Genome Project (HGP) was to generate a highly accurate reference sequence for the human genome. This sequence is now essentially complete and is available in its entirety as a reference for biomedical researchers. High-throughput genome sequencing has created a fundamental shift in the paradigm for biological research. Whereas gene discovery once drove DNA sequencing, now the sequencing of entire genomes drives gene discovery. As such, it is essential that the scientific community be informed about the accuracy of this reference sequence and of its fidelity to the biological templates from which it was derived.

Box 1

Large-scale sequencing terms for this study

Accuracy The measure of how likely the base pairs in a consensus are to be the correct base call. For a 99.99% accurate DNA sequence, it must contain only one incorrect base per 10,000 bp. Accuracy is also sometimes referred to as the 'quality' of a base pair, because estimated base-pair qualities are assigned by the assembly software when it creates the consensus.

Consensus The final reconstructed DNA sequence built by assembling the sequence reads and generating a consensus base call for each position in the assembly. In the case of a finished clone, there is only one consensus.

Contiguity The measure of how many pieces are contained within the assembly. A contiguous assembly would typically have multiple overlapping sequence reads the entire length of the consensus. The finishing rules allow a consensus in more than one piece to be called contiguous (no gaps) in certain difficult situations if the break point is annotated in the database entry.

Fidelity The fidelity of a consensus is how similar the consensus is to the underlying biological template from which the sequence reads were derived. Fidelity for a genome at the single base-pair level is difficult to measure without identifying and sequencing a different clone from the same position on the same chromosome and examining the difference between the sequences. In this study we evaluated the fidelity of a sequence in reference to the large-insert clone from which the sequence was derived, not the genomic template.

Finishing The process of collecting data, performing computational manipulation to a data set to convert a shotgun assembly into a single high-quality contiguous DNA sequence, and verifying the fidelity of the consensus.

World standards for sequence fidelity (known as the Bermuda Standards) were established at the meeting of HGP principal investigators in 1997 (<http://www.gene.ucl.ac.uk/hugo/bermuda2.htm>). These standards stated that finished sequence should contain less than one error per 10,000 DNA bases (99.99% accuracy), and that the sequence should be contiguous (without gaps). Compliance with the base-pair (bp) accuracy standard was measured by error probability assessments generated by DNA base-calling software¹⁻³ and by examining discrepancies between overlapping clone sequences. Compliance with the contiguity standard was an internal measurement based on each centre's complex sequence-finishing methodology. Over the course of the project, additional standards were created to ensure sequence fidelity (<http://www.genome.wustl.edu/Overview/g16stand.php>).

Although more than 2.8 billion base pairs of unique finished sequence has been generated by the sequencing centres comprising the International Human Genome Sequencing Consortium (IHGSC), until the present study was performed fewer than 5,000,000 bp of this sequence has been verified independently for compliance with the finishing standard⁴. Finished chromosome sequence papers have now been published for 9 of the 24 human chromosomes⁵⁻¹³, with most of these papers estimating that the chromosomal sequence exceeds the 99.99% accuracy measure. To provide a more uniform picture of the finished sequence quality of the human genome, the National Human Genome Research Institute (NHGRI) solicited us to perform a detailed evaluation of the DNA sequence data that was generated for the HGP by seven of the IHGSC centres. We examined more than 34 megabases (Mb) of sequence data for accuracy, contiguity and fidelity (see Box 1), and participated in a computational data exchange with the Wellcome Trust Sanger Institute. This paper contains the results of our analysis of the quality of finished sequence data deposited by these centres in the public human genome databases from February 2001 through to July 2002.

Overview and procedure

Our quality assessment of finished human bacterial artificial chromosome (BAC) sequences was conducted in two rounds. For the first round of analysis, we evaluated the finished sequence produced by the three largest NHGRI-funded sequencing centres: the Baylor Human Genome Sequencing Center, the Washington University Genome Sequencing Center and the Whitehead Institute Center for Genome Research. We selected 120 BAC clones (about 6.7 Mb from each centre) from sequence submissions spanning the six-month period from 15 February 2001 through to 15 August 2001. The second round of analysis evaluated the sequence produced by the four smaller sequencing centres that individually

contributed more than 30 Mb to the human genome: the Genome Therapeutics Corporation, the French National Sequencing Center Genoscope, the University of Washington Genome Center and the RIKEN Genomic Sciences Center. We selected 80 BAC clones (about 3.4 Mb from each centre) from these sequencing centres, spanning the 17-month period from 15 February 2001 through to 30 June 2002 (Supplementary Fig. S1 and Table S1).

We sampled clones throughout the two time periods, and adjusted the number of clones that we selected to be a percentage of clones finished by each centre in each month. The sequencing centres provided us with sequencing read data and glycerol stocks of the large-insert clone. In contrast to previous quality assessments⁴, we created a new subclone library for each clone and sequenced this library to 3–4 times coverage in high-quality base pairs. We generated these reads from both ends of sized plasmid subclones, which gave us the ability to evaluate independently the centre's submission regardless of how the original data were generated. We combined our new reads with the original data and then finished the resulting assembly to a high degree of accuracy, performing directed sequencing reactions from the large-insert clone when necessary. These directed reads included reactions performed with alternative chemistries such as dGTP and Invitrogen sequencing enhancers. All finishing quality analysis was performed using the Phred/Phrap/Consed³ pipeline. We then compared our 'gold standard' consensus to the original submitted consensus and then verified and classified any discrepancies. For each discrepancy, we counted the number of error events and base-pair errors and as necessary classified the error as a significant error or misassembly (see Box 2). We counted an error only if original data generated by the submitting centre supported the correct consensus; in this way, we avoided classifying any large-insert clone growth variations as sequencing errors.

Accuracy results and base-pair errors

Our analysis indicates that all of the sequencing centres surveyed met the standards for 99.99% accuracy over the time period studied. Figure 1 shows the plot of the error events and the base-pair errors for each clone that we assessed. These are plotted as rates, normalized per 10 kilobases (kb) over the length of each clone, and include all incorrect base pairs. Most (184 out of 197) of the clones have less

than 1 bp error per 10 kb, with 59 of the clones having no identified errors. Twelve of the thirteen remaining clones exceed the 1 bp per 10 kb standard owing to significant errors. Disregarding the significant errors, only 1 of the 197 clones exceeded the target error rate because of base-pair errors alone. Cumulative error results for each of the rounds are shown in Table 1. The individual centre base-pair error rates ranged from 1 in 25,420 bp to 1 in 154,479 bp, and significant error rates ranged from none found to 1 in 1.2 Mb (Supplementary Table S2).

The vast majority of error events found in the finished human BAC sequences affected a single base pair in the consensus sequence (411 out of 466, 88.2%). Roughly half (48%) of these errors were single base-pair substitutions, with the remainder (52%) being single base-pair insertions or deletions. The substitution errors were primarily miscalled bases in regions of low quality. However, there are many positions where a miscalled base was incorporated into the consensus sequence despite the presence of multiple high-quality reads with the proper base calls; these are obvious finishing errors. Most of these errors occurred where a single discrepant subclone at that position was given a high-quality score by the base-calling algorithm and the miscalled base was incorporated into the consensus sequence. Additionally, we identified 42 (9% of error events) multiple base-pair insertions, deletions and substitutions of less than 20 bp, most of which were clone mutations in a single subclone that were erroneously included in the consensus.

Box 2

Analysis terms for this study

Base-pair errors The number of base-pair changes between our 'gold standard' consensus and the original submitted sequence.

Error events A count of the number of positions of change in the consensus discovered in the quality assessment process; a contiguous insertion, deletion or erroneous run of multiple base pairs is counted as a single error event because the multiple base-pair errors probably arose from a single process error.

Misassembly A rearrangement or deletion of the consensus caused by the incorrect joining of two similar pieces of sequence that are geographically separated in the true consensus.

Significant error A single error that causes at least 50 contiguous base pairs to be incorrect in the submitted consensus versus our gold standard consensus.

```
CGTGCCGGGGCTAATTATTGGCAAAACGAGCTCTTGTGTAAACATTGAT
|||||
CGTGCCGGGACTAATTATTGGCAAAA*****TTGTAAACATTGAT
```

One error event, 1-bp error

One error event, 10-bp errors

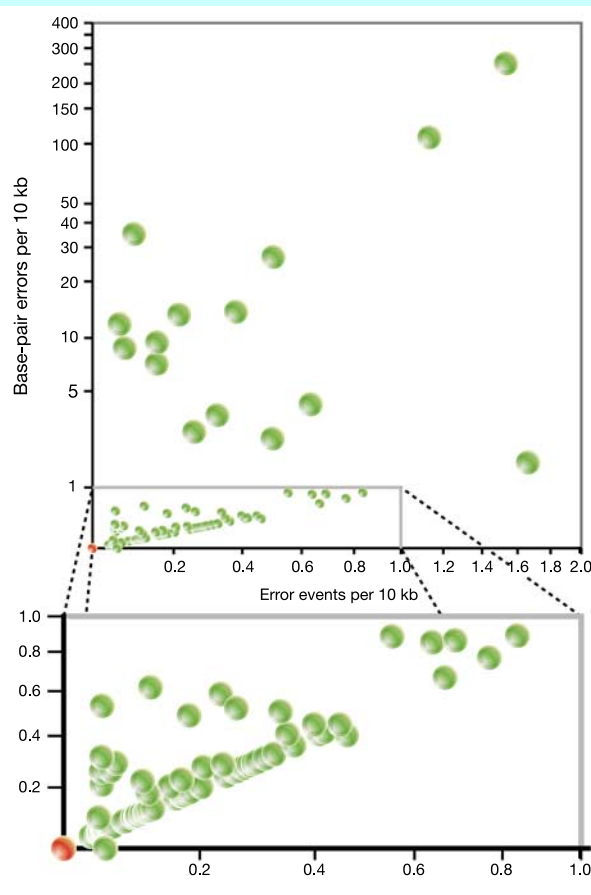


Figure 1 A plot of the error events per 10 kb versus the base-pair errors per 10 kb for the clones surveyed. Each green circle represents a different surveyed clone. A detailed view of the boxed area (less than one error event per 10 kb and less than 1-bp error per 10 kb) shows the diagonal distribution of all of the clones containing only single base-pair errors. The red circle indicates 59 clones with no errors.

Table 1 Cumulative results of each of the quality assessment rounds

Analysis results	Round 1	Round 2	Total
Sequence analysed (kb)	20,303	13,887	34,190
Clones analysed	117	80	197
Error events	183	283	466
Substitution events	73	135	208
Insertion/deletion events	110	148	258
Error event rate (bp)	1/110,948	1/49,069	1/73,369
Base-pair errors*	255	381	636
Base-pair error rate	1/79,621	1/36,448	1/53,758
Significant errors†	5	8	13
Significant error rate (bp)	1/4,060,688	1/1,735,828	1/2,630,005

* Does not include significant insertions, deletions or rearrangements of sequence.

† Insertions or deletions of greater than 50 bp or significant rearrangements of sequence.

Significant errors

We found a significant error in 12 out of the 197 (6.1%) BAC clones that we analysed. There were 13 total significant errors in these clones (2.8% of the total error events). Most of these were identifiable as potential problems from the initial assembly of only the contributing centre's data set. We found large consensus deletions that were derived from deleted subclone templates or polymerase chain reaction (PCR) amplified products. Long stretches of sequence were also deleted as a result of incorrect joins made in repetitive regions, through which sequencing was difficult, and joins were based on minimal sequence overlap. The distribution of these sequence areas that were more difficult to sequence varies across the human genome¹², and consequently, we did not survey difficult clones from every centre.

Potential error-prone finishing techniques

In the course of this quality assessment we identified finishing techniques that in some cases directly contributed to consensus errors that were not corrected before submission by the centre. A large number of the single base-pair-deletion errors were the result of G + C compressions from dye-primer chemistry (now phased out of use in most centres) or dGTP chemistry (a chemistry for difficult-to-sequence regions), or from A or T base drop-out errors on the Megabace platform. Some of the larger deletions in simple sequence regions were from PCR-generated templates or from single subclones that had deleted a portion of the repeat copies. Clones consisting of mostly single-direction M13 reads had more serious assembly issues in repetitive areas. Higher assembly stringencies would have reduced greatly the number of incorrect joins and improved the overall accuracy for the identified misassembled repeat structures.

Computational quality assessment of two contributors

In addition to the quality assessments detailed in this paper, the Wellcome Trust Sanger Institute and the Joint Genome Institute/Stanford Human Genome Center exchanged 38 finished clones over the same time period as round one in this study. These two centres examined only trace data and built new assemblies to compare to the submitted assemblies, and they did not add additional sequencing data. Suspected errors were verified by the original submitting centre. This study found that, for these two centres, there was on average 1 bp error per 651,000 bp and one potential significant error in 11.1 Mb. Together these centres contributed about 39% of the human genome sequence. Although this analysis is not directly comparable with our more detailed study—because computational analysis alone is unable to detect all of the errors found with additional sequencing (see Supplementary text)—this provides a reviewed estimate of error rates for these two centres.

Quality of the finished human genome

We believe that the quality evaluation methodology outlined in this

paper provides a uniform framework to evaluate sequence produced by the disparate finishing systems used by the IHGSC sequencing centres in relation to the standards for finished sequence quality. Of the 197 clones analysed, we found that 182 (92.4%) significantly exceed the 99.99% accuracy standard, on the basis of a calculation of base-pair errors per 10 kb (Fig. 1). If the sampled data set is applicable to the entire genome, we can conclude that the base-pair accuracy standards have been exceeded tenfold, as there is less than 1 bp error per 100,000 bp of finished sequence. If we normalize for the relative amounts of sequence contributed by each centre, we should expect to find on average seven error events with nine incorrect bases per 1 Mb and one significant error per 6 Mb.

We believe that caution should be exercised in extrapolating our data beyond the specific regions of the genome that were surveyed in our study. This quality assessment is an evaluation of process, as it was based on a methodology of sampling sequence production over time, not sampling uniformly from the finished product. As such, our results are a reflection of the finishing methodologies used by the centres for the time period evaluated in our study, and these methodologies were subject to continuous improvement. For the centres investigated in round one, we sampled from a single production period, whereas clones submitted early and late in the HGP were not sampled; these clones are more likely to contain a higher error count. Along with improvements to knowledge and technology, the goals of the overall HGP changed over the course of the project, and the quality threshold used by the sequencing centres fluctuated in response to these production goals. In addition, our thorough quality evaluation methodology (which included additional shotgun sequencing) was applied only to sequencing centres contributing 55% of the total human sequence, with an additional 39% assessed by computational evaluation alone. No sequence was surveyed from the 6% of the genome finished by many smaller contributors.

As a result of differences in the application of finishing methodologies, the most significant factor correlated to sequence quality is centre-to-centre variation (Supplementary Table S2). The sequencing centres had different thresholds at which they determined a given region to be 'finished', and some centres intentionally exceeded the required quality levels (Table S2). The stringency of the contiguity threshold (the cause of most of the significant errors identified) applied by each group was the result of an admixture of production pressures, 'regional' complexity of their genomic territory, personnel experience in addressing the variations in finishing difficulty, and degree of communication and standardization of the group with the larger HGP community. The nature of the HGP as a pilot project for large-scale genomic sequencing makes it difficult to describe the quality of the human genome sequence as a singular entity, although this evaluation has provided valuable insights into the process of producing a complete, complex, finished genome sequence.

Applications to future projects

Well-defined finishing standards specifying targets for accuracy, singular contiguity and fidelity—coupled with descriptions of processes that enable greater compliance with these standards—will enable future genome sequencing projects to generate a more uniform quality product. Continuous sampling of finished sequence for quality evaluation throughout the production process of future genome sequencing projects would better enable global quality statements to be made, and subsequent incorporation of quality assessment feedback into the process could further enhance the quality of the product. Standardizing or minimizing a number of variables—genome source, cloning and library construction platforms, hierarchical sequencing strategies, definitions of finished product—will help further. Such procedures will enhance not only verification ability on a clone or regional basis, but will also be tremendously helpful in solving recalcitrant problems such as the resolution of large duplicated genomic structures. As new genome-sequencing techniques emerge over the course of a sequencing project (for example, cloning vectors, sequence chemistries, detection platforms, finishing techniques), a centralized quality-control centre could serve as a resource for evaluating the technique's relative ability to ensure fidelity with the genomic sequence, rather than each centre independently examining and evaluating all new technologies. In this capacity the quality-control centre would serve as a distributor of reviews and test performance reports for technological developments, which would allow all sequencing centres equal access to information about these techniques. A central trace data repository, such as the NCBI trace archive, is a positive step towards making all raw sequencing trace data available, but also storing the final assemblies would enable central coordination of gap-closing efforts and allow centres to concentrate on the finishing problems that they have developed pipelines to address, instead

of expecting each centre to apply these complicated techniques to an equal standard. □

Received 24 October 2003; accepted 26 January 2004; doi:10.1038/nature02390.

1. Ewing, B. & Green, P. Base-calling of automated sequencing traces using *Phred*. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
2. Ewing, B., Hillier, L., Wendl, M. & Green, P. Base-calling of automated sequence traces using *Phred*. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
3. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
4. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9**, 1–4 (1999).
5. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
8. Helig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
9. Hillier, L. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
10. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
11. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
12. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428**, 493–521 (2004).
14. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the participating centres for providing clone stocks and sequence data sets, and for their feedback about our quality assessment process. We also thank C. Lloyd and C. Bagguley for their detailed computational assessment of our finished sequence.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S. (jeremy@shgc.stanford.edu) or R.M.M. (myers@shgc.stanford.edu).

DNA sequence and analysis of human chromosome 9

S. J. Humphray¹, K. Oliver¹, A. R. Hunt¹, R. W. Plumb¹, J. E. Loveland¹, K. L. Howe¹, T. D. Andrews¹, S. Searle¹, S. E. Hunt¹, C. E. Scott¹, M. C. Jones¹, R. Ainscough¹, J. P. Almeida¹, K. D. Ambrose¹, R. I. S. Ashwell¹, A. K. Babbage¹, S. Babbage¹, C. L. Bagguley¹, J. Bailey¹, R. Banerjee¹, D. J. Barker¹, K. F. Barlow¹, K. Bates¹, H. Beasley¹, O. Beasley¹, C. P. Bird¹, S. Bray-Allen¹, A. J. Brown¹, J. Y. Brown¹, D. Burford¹, W. Burrill¹, J. Burton¹, C. Carder¹, N. P. Carter¹, J. C. Chapman¹, Y. Chen², G. Clarke¹, S. Y. Clark¹, C. M. Clee¹, S. Clegg¹, R. E. Collier¹, N. Corby¹, M. Crosier³, A. T. Cummings¹, J. Davies¹, P. Dhami¹, M. Dunn¹, I. Dutta¹, L. W. Dyer¹, M. E. Earthrow¹, L. Faulkner¹, C. J. Fleming¹, A. Frankish¹, J. A. Frankland¹, L. French¹, D. G. Fricker¹, P. Garner¹, J. Garnett¹, J. Ghor¹, J. G. R. Gilbert¹, C. Glison¹, D. V. Grafham¹, S. Gribble¹, C. Griffiths¹, S. Griffiths-Jones¹, R. Grocock¹, J. Guy³, R. E. Hall¹, S. Hammond¹, J. L. Harley¹, E. S. I. Harrison¹, E. A. Hart¹, P. D. Heath¹, C. D. Henderson¹, B. L. Hopkins¹, P. J. Howard¹, P. J. Howden¹, E. Huckle¹, C. Johnson¹, D. Johnson¹, A. A. Joy¹, M. Kay¹, S. Keenan¹, J. K. Kershaw¹, A. M. Kimberley¹, A. King¹, A. Knights¹, G. K. Laird¹, C. Langford¹, S. Lawlor¹, D. A. Leongamornlert¹, M. Leversha¹, C. Lloyd¹, D. M. Lloyd¹, J. Lovell¹, S. Martin¹, M. Mashregi-Mohammadi¹, L. Matthews¹, S. McLaren¹, K. E. McLay¹, A. McMurray¹, S. Milne¹, T. Nickerson¹, J. Nisbett⁴, G. Nordsiek⁴, A. V. Pearce¹, A. I. Peck¹, K. M. Porter¹, R. Pandian¹, S. Pelan¹, B. Phillimore¹, S. Povey⁵, Y. Ramsey¹, V. Rand¹, M. Scharfe⁴, H. K. Sehra¹, R. Shownkeen¹, S. K. Sims¹, C. D. Skuce¹, M. Smith¹, C. A. Steward¹, D. Swarbreck¹, N. Sycamore¹, J. Tester¹, A. Thorpe¹, A. Tracey¹, A. Tromans¹, D. W. Thomas¹, M. Wall¹, J. M. Wallis¹, A. P. West¹, S. L. Whitehead¹, D. L. Willey¹, S. A. Williams¹, L. Wilming¹, P. W. Wray¹, J. L. Young¹, J. L. Ashurst¹, A. Coulson¹, H. Blöcker⁴, R. Durbin¹, J. E. Sulston¹, T. Hubbard¹, M. J. Jackson³, D. R. Bentley¹, S. Beck¹, J. Rogers¹ & I. Dunham¹

¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

²European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

³The Institute of Human Genetics, The International Centre for Life, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 3BZ, UK

⁴German Research Centre for Biotechnology (GFB), Department of Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany

⁵HUGO Gene Nomenclature Committee, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Chromosome 9 is highly structurally polymorphic. It contains the largest autosomal block of heterochromatin, which is heteromorphic in 6–8% of humans, whereas pericentric inversions occur in more than 1% of the population. The finished euchromatic sequence of chromosome 9 comprises 109,044,351 base pairs and represents >99.6% of the region. Analysis of the sequence reveals many intra- and interchromosomal duplications, including segmental duplications adjacent to both the centromere and the large heterochromatic block. We have annotated 1,149 genes, including genes implicated in male-to-female sex reversal, cancer and neurodegenerative disease, and 426 pseudogenes. The chromosome contains the largest interferon gene cluster in the human genome. There is also a region of exceptionally high gene and G + C content including genes paralogous to those in the major histocompatibility complex. We have also detected recently duplicated genes that exhibit different rates of sequence divergence, presumably reflecting natural selection.

With the completion of the reference sequence for the human genome, the focus now is on detailed analysis to produce a precise and comprehensive depiction of the genome and to identify biologically important features. We present here the highly accurate finished sequence (>99.99%) and analysis of chromosome 9, adding to the completed individual chromosome sequences^{2–8}. In accordance with the goals of the Human Genome Project, our primary aim was to sequence and analyse the euchromatic or gene-containing region of this sub-metacentric chromosome. However, we have also mapped and sequenced substantial portions of the pericentromeric segmentally duplicated regions, which constitute approximately 7% of the chromosome.

Genomic sequence and landscape

A physical map of six contigs spanning the euchromatic part of chromosome 9 (Table 1) was assembled using restriction enzyme fingerprinting and marker content analysis of clones identified by screening up to 90 genomic equivalents of bacterial, P1-derived and yeast artificial chromosome (BAC, PAC and YAC) cosmid and fosmid clone libraries⁹ (Supplementary Table S1). A total of 925 minimally overlapping clones were selected from the map and sequenced (Supplementary Table S2). The latest sequence assembly and the versions analysed here are available at <http://www.sanger.ac.uk/HGP/Chr9>. The features identified in our analysis are shown

in Fig. 1 (rollfold) (for a more detailed view see Supplementary Fig. S1). The sequence of the short arm is contiguous. It contains the 9pter (TTAGGG)*n* telomeric repeat, obtained using YACs containing the captured telomere (H. Riethman, personal communication), and copies of the pericentromeric duplicated sequences. On the long arm there are four small gaps within an 8-megabase (Mb) subtelomeric region (9q34.1–34.3, 128.6–136.5 Mb). The total extent of these gaps (determined by fluorescent *in situ* hybridization of flanking clones to DNA fibres) is <300 kilobases (kb). The absence of clones representing these gaps is probably a consequence of the exceptionally high G+C content in this region (see below). Similar subtelomeric unclonable gaps in (G+C)-rich regions have been seen previously^{2–4}. The most telomeric sequence obtained at 9qter is contiguous with the shortest allelic variant of the subtelomeric repeat. The proximal end of the sequence of the long arm extends into representative blocks of the segmentally duplicated pericentromeric repeats.

A total of 46.15% of chromosome 9 is repeat, similar to previous reports on other chromosomes (for a detailed breakdown of the repeat content see Supplementary Table S3). The G+C content along the chromosome is 41.4% but fluctuates widely, with (G+C)-rich regions correlating with a higher density of both exons and short interspersed nucleotide elements (SINEs). An 8-Mb subtelomeric region of the long arm (in 9q34) has a very high G+C

content (54.2%). Part of this region (near the telomere, 134.7–135.6 Mb) is exceptional, with a G+C content of 59%. In contrast the sequence from 9.0 to 11.0 Mb has a G+C content of 35.1% and a SINE content of just 6.6%.

The completeness and quality of the final sequence assembly was verified by alignment to fosmid and BAC end sequences (<http://genome.cse.ucsc.edu/>) and by comparison with independent genetic and gene map information. We accounted for all of the 603 genes assigned to chromosome 9 in the RefSeq database. All known chromosome 9 markers from three genetic maps^{10–12} were identified in the finished sequence. Comparison of marker order in the sequence with the 193 markers placed on the deCODE genetic map (Fig. 2) identified just one discordant marker placement, D9S1786 relative to D9S1851. On re-examining the deCODE map, these two markers were barely resolved by the genetic data. We conclude that the order of these two markers is as determined in the finished sequence.

Comparison of the genetic map with the finished sequence also enabled us to assess recombination rate along the chromosome (Fig. 2). On the basis of the deCODE map, the mean sex-averaged recombination rate of chromosome 9 is 1.33 cM Mb⁻¹. With the benefit of the finished sequence, we were able to identify a previously unrecognized region of high recombination (as previously defined¹³) between D9S164 and D9S1826 in the subtelomeric region 9q34 (recombination rate 4.5 cM Mb⁻¹).

Gene index

Gene structures were manually annotated in the finished sequence on the basis of computational analysis using predictive algorithms and supporting evidence from expressed sequence tags (ESTs), complementary DNAs and proteins. A total of 1,575 structures, including 1,149 genes and 426 pseudogenes, were categorized as described previously⁴ and are listed in Table 2. The average gene density of the chromosome is 10.5 genes Mb⁻¹, close to the genome average (Supplementary Table S4), but ranges from 3 genes Mb⁻¹ in 9p23 to 22 genes Mb⁻¹ in the subtelomeric region 9q34. Part of the subtelomeric region (134.7–135.6 Mb) is exceptional, with a gene density of 66 genes Mb⁻¹, which is associated with a very high G+C content of 59%. Annotated genes (excluding pseudogenes) cover 47% of the sequence, whereas exons occupy 2.5% of the total and have a mean length of 326 base pairs (bp). The longest exon (10,135 bp) lies in a gene of unknown function (KIAA1958), whereas the longest intron spans 582 kb in *TRPM3*. The longest gene is protein tyrosine phosphatase receptor type D (*PTPRD*), a gene implicated in synaptic plasticity and new memory acquisition¹⁴ that spans 2.3 Mb and contains 46 exons. This gene, which occurs in the gene-poor 9p23 region, is one of the largest genes found in the human genome so far, being comparable in size to the 2.3-Mb dystrophin gene on Xp12. We identified 432 CpG islands (see Supplementary Methods) within a window of 5 kb upstream and 1 kb downstream of the 613 'known' genes (that is,

genes that match a known messenger RNA or protein sequence); thus 70% of these genes have an associated CpG island.

Alternative splicing of genes generates multiple transcripts, leading to diversity of protein structures. On chromosome 9 the mean number of transcripts annotated per gene is 3.09. In some cases alternative splicing was extensive; for example, we annotated 26 different transcripts in the *CIZ1* gene, which encodes a nuclear protein potentially involved in brain tumorigenesis¹⁵. Eight of these transcripts have open reading frames (ORFs) encoding different protein isoforms, two of which are partial and do not contain the zinc finger domain.

MicroRNA (miRNA) genes encode RNA products of around 22 nucleotides (<http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml>) and have been implicated in gene regulation. We have detected 14 miRNA genes on chromosome 9 including two clusters of three genes in 9q22. All 14 miRNAs are conserved in mouse with respect to gene order and orientation, and the two human clusters have counterparts on mouse chromosome 13. We also identified eight transfer RNA genes using tRNAscan-SE distributed along the chromosome.

Comparative analysis was used as an independent measure of the completeness of gene annotation of the protein-coding genes. We identified 4,190 evolutionarily conserved regions (ECRs; see Supplementary Methods) that are conserved in the sequence of human chromosome 9 and the genomic sequence of five other species (mouse, rat, *Fugu*, *Tetraodon* and zebrafish; see Supplementary Table S5). A total of 4,091 of the ECRs overlapped 4,516 exons that had previously been annotated, whereas no evidence was found for the existence of coding transcripts on either DNA strand of the human sequence for 99 of the ECRs. On the basis of this observation we conclude that the annotation of the protein-coding exons is at least 97.9% (=4,516/(4,516 + 99)) complete. There was a notable cluster of 22 ECRs in the introns of two flanking genes, *MAPKAP1* and *PBX3* (123.7–124.3 Mb, see Supplementary Fig. S2a, b), and these ECRs are therefore good candidates to test for possible regulatory function.

Sequence duplication

Gene duplication events give rise to multigene families. In some cases, subsequent evolutionary divergence of duplicated genes gives rise to new or altered protein function¹⁶. We searched the sequence for gene clusters likely to have arisen by duplication using BLASTP (as previously defined⁸) and on the basis of known protein domains using InterProScan (<http://www.ebi.ac.uk/interpro/scan.html>) (Supplementary Table S6). The BLASTP analysis identified 31 clusters containing 98 genes (Supplementary Fig. S1 and Table S7). There are 16 interferon type 1 genes in the human genome, all of which are on chromosome 9. Fifteen are clustered in a single 364-kb region (21.0–21.4 Mb) within a genomic duplication (Fig. 3a; see also Supplementary Fig. S1). This gene cluster is also present in the orthologous mouse sequence, indicating that some expansion of the type 1 interferons occurred before divergence of human and mouse from a common mammalian ancestor. The top-ranking

Table 1 Sequence contigs on chromosome 9

Contig*	Size (Mb)	Gap size (kb)
AL928970–BX005214	39.4	–
Pericentromere/centromere	–	ND
AL353608–AL360004	62.0	–
Gap 1	–	50
AL354898–AL591386	3.8	–
Gap 2	–	200
AL354796–AL683798	0.2	–
Gap 3	–	13
AL669970–AL138781	1.8	–
Gap 4	–	30
AL603784–AL954642	1.9	–
Total	109.1	293
Total euchromatic region	109.4	–

*Contigs are indicated by the first and last sequence accession. ND, not determined.

Figure 1 Chromosome 9 sequence features (see rollfold). Tracks from top to bottom are: (1) sequence scale (Mb); (2) coverage of chromosome 9 sequence (black) and gaps (grey); (3) synteny to mouse (top track) and rat (bottom track) chromosomes, with chromosomes colour-coded and coordinate range (Mb) indicated (Un/random indicate that there is no current chromosome location for the homologous mouse or rat sequences); (4) position of predicted CpG islands (brown); (5) location of ECRs showing sequence homology to *Fugu* (blue), zebrafish (dark blue) and *Tetraodon* (dark pink); (6) placement of 'known' (dark blue) and 'novel coding sequence' (black) annotated gene structures (official gene symbols used when available). Owing to space restrictions this figure represents an abbreviated set of features and we therefore recommend downloading Supplementary Fig. S1 to follow the text accurately.

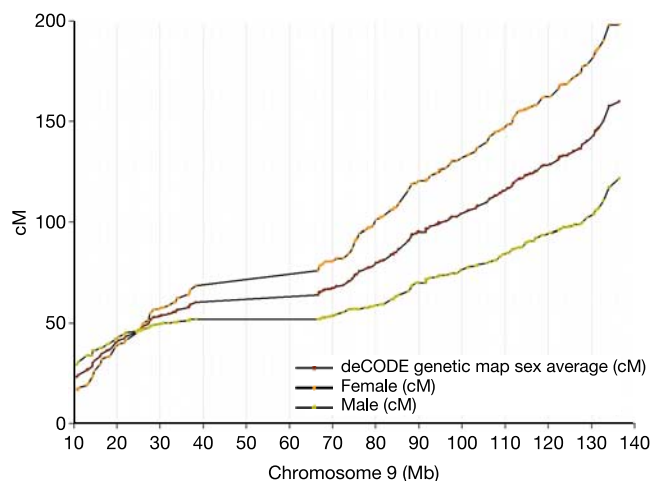


Figure 2 The relationship between the physical and deCODE¹⁰ genetic distance. The physical location of each genetic marker is shown on the female, male and sex-averaged genetic maps.

InterProScan hits (listed in Supplementary Table S6) show good agreement with the clustering by BLASTP.

Some gene duplications are observed in human but not mouse, indicating that they occurred since the divergence of these two species. On the basis of this criterion we identified 18 sets of recently duplicated genes on chromosome 9 (both intra- and interchromosomal duplications; see Methods and Supplementary Table S8). These included two relaxin genes, five forkhead genes, two distinct sets of olfactory receptor genes, a pair of orosomucoid (ORM) genes, and the 13 interferon A genes within the type 1 cluster. Note that in the case of the interferon genes, there is evidence of expansion of these genes since the human–mouse divergence as well as duplication earlier in mammalian evolution (described above). The accumulation of synonymous changes between duplicated genes gives an indication of the time that has elapsed since the duplication event, on the assumption that there is no selection acting on such changes. The rate of synonymous base substitutions between duplicated sequences is given by the parameter K_s (ref. 17). For example, comparison of the olfactory receptor genes *OR13C3* and *OR13C4* gives a high K_s value (0.372), suggesting that this duplication is less recent than the duplication that gave rise to the forkhead genes *FOXD4L2* and *FOXD4L4*, which have a much lower K_s value (0.0134) (Supplementary Table S8).

Evolutionary divergence of genes after duplication may arise as a result of natural selection acting on new mutations at either locus. The accumulation of non-synonymous changes (given by the parameter K_a (ref. 17)) relative to the rate of synonymous change (K_s) gives an indication of selection expressed as the ratio K_a/K_s . Comparison of the two relaxin genes (*RLN1* and *RLN2*) gives a K_s of

0.119 and a K_a of 0.0999, and thus one of the highest K_a/K_s ratios of 0.84 among the recent duplications. By contrast the olfactory receptors *OR13C3* and *OR13C4*, despite being the result of a much less recent duplication, appear to be more conserved ($K_s = 0.372$, $K_a = 0.0843$, $K_a/K_s = 0.23$) compared with the relaxin genes. Note, however, that these results are based on small numbers of nucleotide substitutions (Supplementary Table S8).

Phylogenetic reconstruction using the relaxin genes along with additional sequences from chimpanzee (Z27225, Z27245), rhesus macaque¹⁸, lemur (AF317625) and pig (J02792) demonstrated that the relaxin duplication occurred after the divergence of lemurs from other primates but at least before divergence of higher primates from Old World monkeys. A likelihood ratio test indicated that positive selection has acted on the relaxin sequences since duplication ($P < 0.001$) (see Methods and ref. 19). There is evidence for *RLN2* being selectively expressed in the ovary during pregnancy where it acts to facilitate parturition²⁰. There is no evidence that *RLN1* is expressed in the ovaries but it may be expressed in other tissues²¹. Hence, any evolved differences in function of the duplicated relaxin genes may represent an example of tissue-specific sub-functionalization. The differing levels of sequence similarity seen across the genes (Fig. 4) may reflect these alternative functions. Further analysis of these genes may uncover functional explanations for this positive selection.

Extensive genomic segmental duplications were reported previously for the draft sequence^{22,23}, with intra- and interchromosomal repeats estimated to involve from 5.5% to 7.1% and 3.6% to 4.7% of chromosome 9, respectively, the fourth highest level among the autosomes. The centromere of chromosome 9 is flanked by heterochromatic C bands rich in satellite sequences. Heteromorphism of 9qh is common and pericentric inversions occur in 1% of the population²⁴. We generated over 7 Mb of finished sequence from the pericentromeric regions, which currently lie in 33 contigs of mean size 262 kb. This sequence has been included in an analysis of intra- (Fig. 3a) and interchromosomal (Fig. 3b) duplications based on the method of ref. 23.

The pericentromeric contigs contain no unique DNA, with all sequences sharing high (>95%) sequence identity to sequences within chromosome 9 or elsewhere. Many pericentromeric sequences are duplicated in the region and we identified large blocks of duplications within and between the two pericentromeric regions at 9p11–12 and 9q11–12/13 with an average sequence identity of >97% (Fig. 3a). There are also matches to pericentromeric and non-pericentromeric regions of other chromosomes; for example, duplication of a set of genes from the ancestral chromosome fusion site at 2q13 (ref. 25) including an active *FOXD4* gene at 9p24 and copies in the pericentromeric sequences (Fig. 3a, b). These characteristics are consistent with formation of intrachromosomal and interchromosomal duplications^{26,27}.

Despite the duplicated nature of the chromosome 9 pericentromeric sequence, 138 gene features were annotated within it (summarized in Supplementary Table S9). There are only two known

Table 2 Annotation summary

Class*	Gene count	Total length in bp (% coverage)	Mean gene length (bp)	Mean exon length	Number of exons
Known genes	613	41,424,864 (37.9)	68,440	322	6,628
Novel genes	147	4,707,061 (4.3)	32,024	474	605
Novel transcripts	176	4,655,574 (4.3)	26,974	296	683
Putative genes	213	1,912,134 (1.7)	8,980	244	488
Total genes	1,149	51,435,292 (47.0)	46,407	326	8,404
Processed pseudogenes	397	459,658 (0.4)	1,157	709	475
Unprocessed pseudogenes	29	405,851 (0.4)	13,994	165	216
Total structures	1,575	52,022,232 (48.0)	34,405	342	9,095

*Classes follow those laid down in ref. 4 and are based on the longest transcript of each gene; the total number of exons from all alternative transcripts is 14,640. Known genes, identical to known human cDNA or protein and have an entry in Locuslink; novel genes, have an ORF and are identical to spliced human ESTs or have some similarity to other cDNA/proteins (all species); novel transcripts, similar to novel genes but have ambiguous ORFs or multiple evidence for non-coding RNA; putative genes, with identity to 1–3 spliced ESTs but do not contain evidence for an ORF; pseudogenes, similar to known proteins but contain a frameshift and/or stop codon that disrupts the ORF.

genes, one of which contains an SST1 satellite sequence within its predicted ORF and is only expressed in neoplastic tissue. A total of 120 annotated structures are pseudogenes or transcripts with no associated ORF. The remaining 16 structures include many transcripts specific to either testis or neoplastic tissue. This restricted transcription pattern is consistent with previous analyses of similar regions on chromosomes 10 (ref. 28) and 21 (ref. 29).

This analysis revealed much recent segmental duplication; however, it has also been proposed that 9q may have been involved in a more ancient event, leading to 9q34 containing genes paralogous to the major histocompatibility complex on chromosome 6. 'Block duplication' has recently been proposed as an alternative to 'chromosome/genome duplication' or 'selection-based clustering' for the underlying duplication mechanism(s)³⁰. We have placed 20

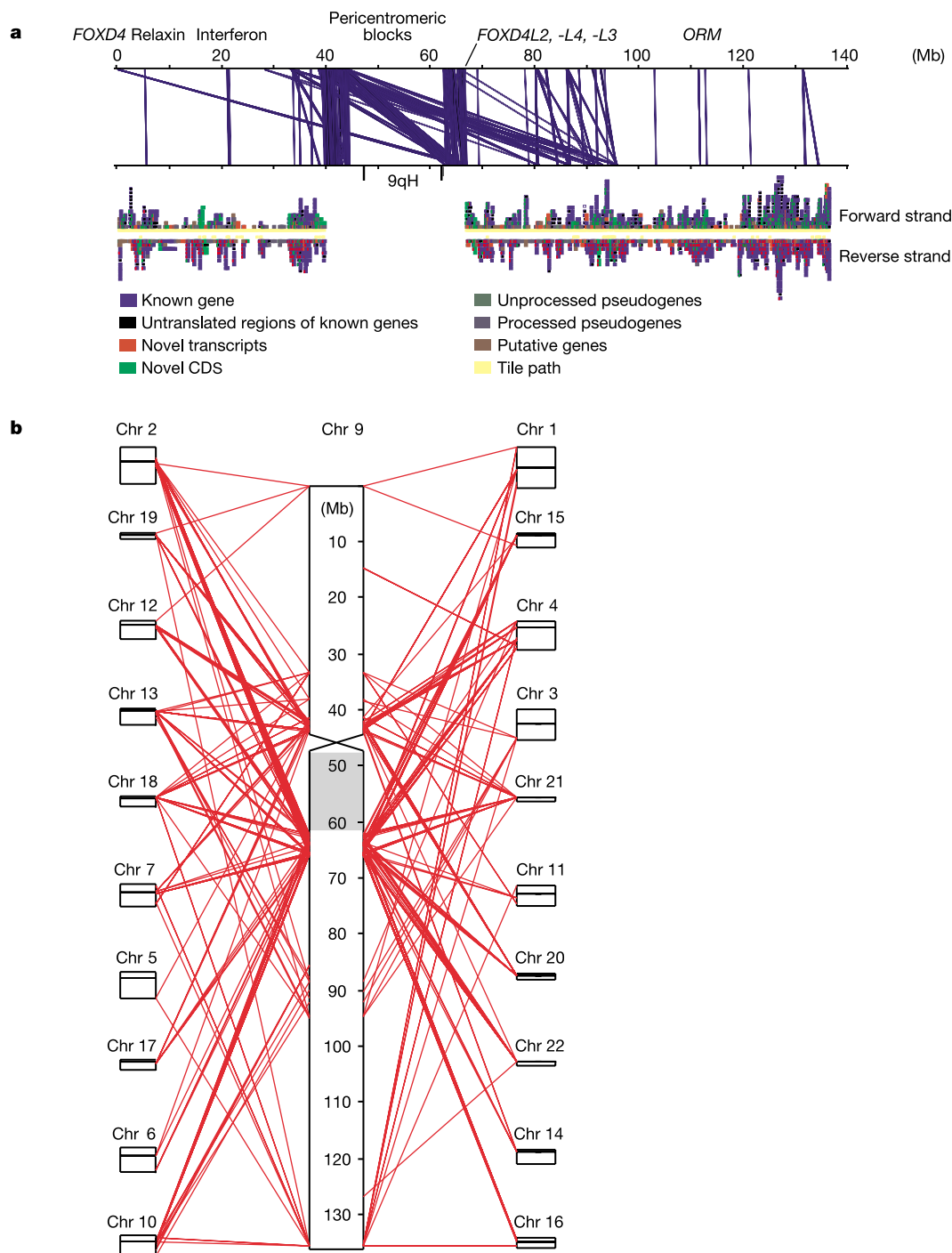


Figure 3 Segmental duplications on chromosome 9. **a**, Intrachromosomal duplications. The top panel shows duplications plotted against the sequence coordinates (matches of >1 kb in length), whereas the bottom panel shows annotated genes on the forward and reverse strand. Some of the gene duplications we identified are also noted. The view is produced using Apollo (<http://www.fruitfly.org/annot/apollo>). **b**, Interchromosomal

duplications. Large duplications (>10 kb) are shown, with chromosome 9 magnified. Other chromosomes are arranged in the order in which they match to chromosome 9. Chromosomes 8, 14, X and Y are not detected at this level. The grey rectangle is heterochromatin (47–62 Mb). Supplementary Fig. S3 shows interchromosomal duplication matches at >1 kb.

novel or previously unlocalized putative major histocompatibility complex gene paralogues within a block of 24 Mb on 9q34 (112.9–136.5 Mb), bringing the total to 31 putative paralogues between this and the corresponding region of chromosome 6 (Supplementary Table S10). Considering an estimated duplication time of around 500 million years ago, it is not surprising that the overall gene order is poorly conserved. Nevertheless, the clustering of these paralogues is highly nonrandom ($P < 0.0001$).

Sequence variation

Inherited sequence variations have an important role in determining variable disease and drug response, and provide the basis for studies of disease association and population genetics³¹. The most abundant form of sequence variation is single nucleotide polymorphisms (SNPs). We have mapped 117,405 SNPs on the sequence (1,073 SNPs Mb⁻¹). A total of 51,122 SNPs (43.5%) were generated during the sequencing by comparing sequence in clone overlaps, whereas the remaining 91,217 were extracted from dbSNP (<http://www.ncbi.nlm.nih.gov/SNP>). The density of mapped random SNPs is clearly higher within the duplicated sequences (Supplementary Fig. S1) at 9p24, those close to the *AQP7* locus (33.5–33.6 Mb) at 9p13, and those flanking the pericentromeric sequences on 9p (38.8–39.4 Mb) and 9q (66.5–66.7 Mb). The elevated SNP density in these regions probably reflects a high rate of misplaced SNPs or paralogous sequence variants²³. After adjusting for bias by removing SNPs generated by directed sequencing, we established *MCART1* (mitochondrial carrier triple repeat 1) as the most polymorphic gene on chromosome 9 with 16 coding SNPs kb⁻¹.

We have also mapped 617,071 sites of sequence divergence between chimpanzee and human from a set of 14 million chimpanzee sequences, released by the Broad Institute and Washington University Genome Sequencing Centre. These sequence differences are evenly distributed except in three regions: two close to 9pter and another one in the 9q pericentromeric region at 66.6 Mb (Supplementary Fig. S1). Two of these regions also have a higher human SNP density and are probably caused by mismapping due to duplicated sequences. A total of 0.6% of the unmatched variations (where there is no corresponding human SNP) are in coding regions

(Supplementary Table S11). We assessed selection levels for the chromosome 9 genes by calculating K_a/K_s values compared to chimpanzee. Examples of genes outside duplicated regions that have elevated K_a/K_s values, and may therefore have undergone marked divergence since their common ancestor, include *C9orf37* (K_a/K_s of 2.37) and *IL11RA* (K_a/K_s of 2.12) (Supplementary Table S12).

Medical implications

There are 95 genes on chromosome 9 that are known to be associated with a disease (OMIM and GeneCard). Sixty-eight of these associations have been detected in the sequence (Supplementary Table S13). The remaining 27 disease loci are currently defined by linkage analysis (Supplementary Table S14). The complete sequence and annotation of the chromosome provides a valuable tool for identifying the genes involved in disease as well as a starting point for genetic analysis of multifactorial diseases. The finished sequence has already enabled the identification of several new genes involved in disease. Chorea-acanthocytosis is an autosomal recessive neurodegenerative disorder with neurological symptoms resembling Huntington's chorea. A novel causative gene (*CHAC*; also known as *VPS13A*) involved in protein sorting was identified from the genomic sequence³². *CHAC* is also notable for having 72 exons, the highest number for a chromosome 9 gene. Sequences generated from the autosomal recessive Hereditary Inclusion Body Myopathy (HIBM) critical region were used to identify a 13-exon gene, *GNE*, which has been established as the gene involved in the disorder³³.

There are three genes on 9p24, *DMRT1*, -2 and -3, which code for proteins with a DNA-binding motif (DM domain) and are expressed in testis. DM domain genes control sex determination in many species, including *Drosophila* and *Caenorhabditis elegans*³⁴. High *DMRT1* expression is necessary for testicular differentiation³⁵, and hemizygous deletion of 9p24 results in haploinsufficiency and human XY sex reversal³⁶. There is also evidence for another gene in the region involved in XY sex reversal, *NR5A1* (ref. 37). The presence of two such loci may reflect extensive conserved synteny between parts of chromosome 9 and the chicken sex-determining chromosome Z³⁸, because 17 of the 24 known chicken Z genes have chromosome 9 orthologues³⁹.

Chromosome 9p21 harbours the tumour suppressor gene *CDKN2A* and is often involved in translocations or deletions in tumour cell lines. Mutation or loss of *CDKN2A* is linked to familial melanoma and other tumours⁴⁰. Reciprocal translocation between chromosomes 9 and 22 leads to generation of a derivative chromosome, the Philadelphia chromosome, in >90% of chronic myelogenous leukaemia and 5–15% of acute B cell lymphoblastic leukaemias⁴¹. The translocation results in fusion of *ABL1* and the chromosome 22 gene *BCR*⁴² to give a chimaeric oncogene with deregulated tyrosine kinase activity⁴³. Inhibition of the BCR-ABL tyrosine kinase is the basis of the anti-cancer activity of imatinib mesylate (Gleevec)⁴⁴. The ability to use genomic information to develop effective new therapies such as this will benefit enormously from finished human chromosome sequences. □

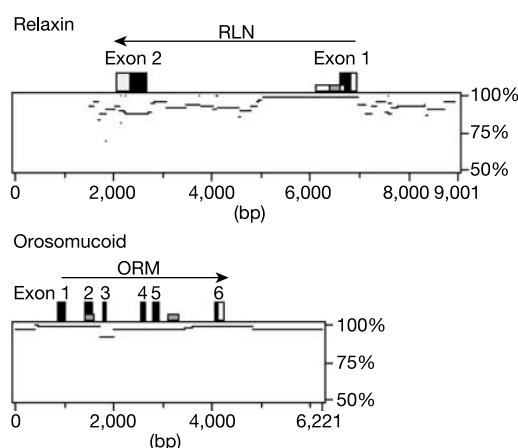


Figure 4 Alignments between duplicated chromosome 9 sequences using MultiPipMaker⁵⁰, indicating percentage identity for each alignment. A genomic duplication event seems to have taken place with a 7,144-bp sequence (5,289,576–5,296,720 bp) and a 7,125-bp region (5,324,724–5,331,849 bp), duplicating the two-exon 5-kb relaxin (RLN) genes. The sequence homology across the duplication shows variation, with the region around the first exon and into the intron matching at 97%, whereas there is a marked decrease in similarity towards exon 2. In comparison, another recently duplicated gene pair, orosomucoid (ORM), duplicated within a 6,212-bp sequence (112,627,806–112,634,018 bp) and 6,220 bp (112,634,577–112,640,797 bp) show more homogeneous sequence alignment.

Methods

Establishing sequence-ready bacterial contigs covering chromosome 9 was a prerequisite for clone-by-clone sequencing, and the methods have been described previously⁹. Briefly, publicly available markers used included 916 ESTs, 763 of which mapped onto the Genemap '98 radiation hybrid map, 229 polymorphic microsatellites from genetic maps, and 226 from RHmaps (Stanford and WI), which were used to screen initially the RPCI-11 BAC library (<http://bacpac.chori.org/hmale11.htm>). These were later augmented with 633 sequence-tagged sites (STSs) derived from single pass sequences of plasmid subclones made from flow-sorted chromosomes. Owing to the similarity in base composition and size of chromosomes 9–12, a pool of samples containing the four chromosomes was used. Fingerprints for positive clones were extracted from humanmap (<http://genome.wustl.edu/projects/human/index.php?pc=1>) or generated in-house⁴⁵, then assembled into contigs in FPC⁴⁶ on the basis of shared *HindIII* fragments and individual clone marker content. Novel STSs were designed from sequences at the ends of each contig

and used in further iterative walking experiments onto an increasing number of different libraries to generate new coverage across gaps. In cases where no bacterial clone coverage could be identified YAC and chromosome-specific cosmid libraries were screened. Extensive checks were performed on each clone before sequencing.

Methods used for shotgun sequencing, finishing strategies, comparative analysis and identification of ECRs, sequence annotation predications and SNP identification are as previously described⁸, and are available as Supplementary Methods.

Gene duplications

A database of manually annotated chromosome 9 genes and Ensembl genes from human (build 15.33, excluding genes from chromosome 9), mouse (build 15.2) and rat (build 15.2) was generated. Homologues of each chromosome 9 gene, excluding pseudogenes, were found using BLASTN⁴⁷, and the genetic distance between the full extent of these sequences was determined¹⁷.

Matches between a chromosome 9 gene and another human sequence were classed as probable recent duplications in cases where the synonymous genetic distance between the human sequences was less than the synonymous genetic distance between the human sequence and the closest mouse or rat sequence. In cases where there was no match to a rodent an arbitrary, synonymous genetic distance cut-off of 0.6 was applied. Likely duplications were disregarded where a substantial disagreement in the number of exons between genes existed or where a match sequence was annotated as a possible pseudogene.

Phylogenetic reconstruction of the relaxin sequences was performed using fastDNAmL⁴⁸. The likelihood ratio test of positive selection was performed following the method of ref. 19 using nested models M1 and M2.

SsaHaSNP⁴⁹ was used to map chimpanzee reads (approximately 14 million) and call single base sequence differences against the human NCBI31 assembly. Those that mapped to chromosome 9 were remapped to the latest Sanger assembly.

Segmental duplication

Various methods have been proposed for the detection of segmental duplications in the human genome^{22,23}. Our method is based on that described by ref. 23. WU-BLASTN⁴⁷ (<http://blast.wustl.edu>) was used to perform all searches.

To detect intrachromosomal duplications, the sequence of chromosome 9 (SANGER_10 assembly) was searched against itself, and 'self' matches (that is, occurring at the same position in the query and target) were discarded. Interchromosomal duplications were detected by searching the chromosome sequence against the non-chromosome-9 parts of the NCBI34 build of the human genome (http://www.ensembl.org/Homo_sapiens/).

Received 22 December 2003; accepted 8 March 2004; doi:10.1038/nature02465.

- Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9**, 1–4 (1999).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Chromosome 21 Mapping and Sequencing Consortium. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
- Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
- Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
- Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
- Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
- Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409**, 942–943 (2001).
- Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
- Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
- Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
- Uetani, N. *et al.* Impaired learning with enhanced hippocampal long-term potentiation in PTPδ-deficient mice. *EMBO J.* **19**, 2775–2785 (2000).
- Warder, D. E. & Keherly, M. J. Ciz1, Cip1 interacting zinc finger protein 1 binds the consensus DNA sequence ARYSR(0-2)YYAC. *J. Biomol. Sci.* **10**, 406–417 (2003).
- Ohno, S., Wolf, U. & Atkin, N. B. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).
- Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
- Crawford, R. J., Hammond, V. E., Roche, P. J., Johnston, P. D. & Tregear, G. W. Structure of rhesus monkey relaxin predicted by analysis of the single-copy rhesus monkey relaxin gene. *J. Mol. Endocrinol.* **3**, 169–174 (1989).
- Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
- Hudson, P. *et al.* Structure of a genomic clone encoding biologically active human relaxin. *Nature* **301**, 628–631 (1983).
- Bathgate, R. A., Samuel, C. S., Burazin, T. C., Gundlach, A. L. & Tregear, G. W. Relaxin: new peptides, receptors and novel actions. *Trends Endocrinol. Metab.* **14**, 207–213 (2003).
- Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

- Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
- Madan, K. & Bobrow, M. Structural variation in chromosome No 9. *Ann. Genet.* **17**, 81–86 (1974).
- Fan, Y., Newman, T., Linardopoulou, E. & Trask, B. J. Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. *Genome Res.* **12**, 1663–1672 (2002).
- Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10**, 2215–2223 (2001).
- Park, J. P., Wojcicki, S. A., Spellman, R. A., Rhodes, C. H. & Mohandas, T. K. Human chromosome 9 pericentric homologies: implications for chromosome 9 heteromorphisms. *Cytogenet. Cell Genet.* **182**, 192–194 (1998).
- Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13**, 159–172 (2003).
- Brun, M. E., Ruault, M., Ventura, M., Roizes, G. & De Sario, A. Juxtacentromeric region of human chromosome 21: a boundary between centromeric heterochromatin and euchromatic chromosome arms. *Genome* **312**, 41–50 (2003).
- Flajnik, M. F. & Kasahara, M. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351–362 (2001).
- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Rampoldi, L. *et al.* A conserved sorting-associated protein is mutant in chorea-acanthocytosis. *Nature Genet.* **28**, 119–120 (2001).
- Eisenberg, I. *et al.* The UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase gene is mutated in recessive hereditary inclusion body myopathy. *Nature Genet.* **29**, 83–87 (2001).
- Raymond, C. S. *et al.* Evidence for evolutionary conservation of sex-determining genes. *Nature* **391**, 691–695 (1998).
- Smith, C. A., McClive, P. J., Western, P. S., Reed, K. J. & Sinclair, A. H. Conservation of a sex-determining gene. *Nature* **402**, 601–602 (1999).
- Shan, Z. *et al.* FISH mapping of the sex-reversal region on human chromosome 9p in two XY females and in primates. *Eur. J. Hum. Genet.* **8**, 167–173 (2000).
- Luo, X., Ikeda, Y. & Parker, K. L. A cell-specific nuclear receptor is essential for adrenal and gonadal development and sexual differentiation. *Cell* **77**, 481–490 (1994).
- Nanda, I. *et al.* Conserved synteny between the chicken Z sex chromosome and human chromosome 9 includes the male regulatory gene DMRT1: a comparative (re)view on avian sex determination. *Cytogenet. Cell Genet.* **89**, 67–78 (2000).
- Schmid, M. *et al.* First report on chicken genes and chromosomes 2000. *Cytogenet. Cell Genet.* **90**, 169–218 (2000).
- Della Torre, G. *et al.* CDKN2A and CDK4 mutation analysis in Italian melanoma-prone families: functional characterization of a novel CDKN2A germ line mutation. *Br. J. Cancer* **85**, 836–844 (2001).
- Wong, S. *et al.* IL-3 receptor signaling is dispensable for BCR-ABL-induced myeloproliferative disease. *Proc. Natl Acad. Sci. USA* **100**, 11630–11635 (2003).
- Chissoe, S. L. *et al.* Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27**, 67–82 (1995).
- Pendergast, A. M., Gishizky, M. L., Havlik, M. H. & Witte, O. N. SH1 domain autophosphorylation of P210 BCR/ABL is required for transformation but not growth factor independence. *Mol. Cell. Biol.* **13**, 1728–1736 (1993).
- Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).
- Humphray, S. J., Knaggs, S. J. & Ragoussis, I. Contiguation of bacterial clones. *Methods Mol. Biol.* **175**, 69–108 (2001).
- Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10**, 41–48 (1994).
- Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
- Schwartz, S. *et al.* PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the *Fugu* and *Tetraodon* sequencing groups, the Mouse Sequencing Consortium, the Zebrafish sequencing groups at the Wellcome Trust Sanger Institute, the Rat Genome project at Baylor College of Medicine, and the Broad Institute and Washington University Genome Sequencing Centre for their early release of the chimpanzee sequences. We also thank H. Riethman for the provision of clones, V. K. Khodiyar, H. M. Wain, E. A. Bruford, M. W. Wright, R. C. Lovering, C. C. Talbot and M. J. Lush from the HUGO Gene Nomenclature Committee for the official gene nomenclature, the German Federal Ministry for Education and Research (BMBF) through DLR for financial support of the GBF group, and the EMBL and Ensembl database teams at the European Bioinformatics Institute. Work at the Institute of Human Genetics was funded by the Wellcome Trust, and work at HUGO was funded by the NIH and the MRC. Work at the Sanger Institute was funded by the Wellcome Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J.H. (sjh@sanger.ac.uk). Accession numbers for the sequences analysed for this paper can be found in Supplementary Fig. S1. All reported DNA sequences have been deposited in EMBL, GenBank or DDBJ.

The DNA sequence and comparative analysis of human chromosome 10

P. Deloukas¹, M. E. Earthrowl¹, D. V. Grafham¹, M. Rubenfield^{2,3}, L. French¹, C. A. Steward¹, S. K. Sims¹, M. C. Jones¹, S. Searle¹, C. Scott¹, K. Howe¹, S. E. Hunt¹, T. D. Andrews¹, J. G. R. Gilbert¹, D. Swarbreck¹, J. L. Ashurst¹, A. Taylor¹, J. Battles², C. P. Bird¹, R. Ainscough¹, J. P. Almeida¹, R. I. S. Ashwell¹, K. D. Ambrose¹, A. K. Babbage¹, C. L. Bagguley¹, J. Bailey¹, R. Banerjee¹, K. Bates¹, H. Beasley¹, S. Bray-Allen¹, A. J. Brown¹, J. Y. Brown¹, D. C. Burford¹, W. Burrill¹, J. Burton¹, P. Cahill², D. Camire², N. P. Carter¹, J. C. Chapman¹, S. Y. Clark¹, G. Clarke¹, C. M. Clee¹, S. Clegg¹, N. Corby¹, A. Coulson¹, P. Dhami¹, I. Dutta¹, M. Dunn¹, L. Faulkner¹, A. Frankish¹, J. A. Frankland¹, P. Garner¹, J. Garnett¹, S. Gribble¹, C. Griffiths¹, R. Grocock¹, E. Gustafson^{2,3}, S. Hammond¹, J. L. Harley¹, E. Hart¹, P. D. Heath¹, T. P. Ho², B. Hopkins¹, J. Horne², P. J. Howden¹, E. Huckle¹, C. Hynds², C. Johnson¹, D. Johnson¹, A. Kana², M. Kay¹, A. M. Kimberley¹, J. K. Kershaw¹, M. Kokkinaki⁴, G. K. Laird¹, S. Lawlor¹, H. M. Lee², D. A. Leongamornlert¹, G. Laird¹, C. Lloyd¹, D. M. Lloyd¹, J. Loveland¹, J. Lovell¹, S. McLaren¹, K. E. McLay¹, A. McMurray¹, M. Mashreghi-Mohammadi¹, L. Matthews¹, S. Milne¹, T. Nickerson¹, M. Nguyen², E. Overton-Larty¹, S. A. Palmer¹, A. V. Pearce¹, A. I. Peck¹, S. Pelan¹, B. Phillimore¹, K. Porter¹, C. M. Rice¹, A. Rogosin^{2,3}, M. T. Ross¹, T. Sarafidou⁴, H. K. Sehra¹, R. Shownkeen¹, C. D. Skuce¹, M. Smith¹, L. Standring², N. Sycamore¹, J. Tester¹, A. Thorpe¹, W. Torcasso², A. Tracey¹, A. Tromans¹, J. Tsolas^{2,3}, M. Wall¹, J. Walsh², H. Wang², K. Weinstock², A. P. West¹, D. L. Willey¹, S. L. Whitehead¹, L. Wilming¹, P. W. Wray¹, L. Young¹, Y. Chen⁵, R. C. Lovering⁶, N. K. Moschonas⁴, R. Siebert⁷, K. Fechtel², D. Bentley¹, R. Durbin¹, T. Hubbard¹, L. Doucette-Stamm^{2,3}, S. Beck¹, D. R. Smith^{2,3} & J. Rogers¹

¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

²Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453, USA

³Agencourt Bioscience Corporation, 100 Cummings Center, Beverly, Massachusetts 01915, USA

⁴Department of Biology, University of Crete & Institute of Molecular Biology and Biotechnology, Foundation of Research and Technology, PO Box 2208, 71409 Heraklion, Crete, Greece

⁵European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

⁶HUGO Gene Nomenclature Committee, Department of Biology, University College London, London NW1 2HE, UK

⁷Institute of Human Genetics, University Hospital Schleswig-Holstein Campus Kiel, Schwanenweg 24, D-24105 Kiel, Germany

The finished sequence of human chromosome 10 comprises a total of 131,666,441 base pairs. It represents 99.4% of the euchromatic DNA and includes one megabase of heterochromatic sequence within the pericentromeric region of the short and long arm of the chromosome. Sequence annotation revealed 1,357 genes, of which 816 are protein coding, and 430 are pseudogenes. We observed widespread occurrence of overlapping coding genes (either strand) and identified 67 antisense transcripts. Our analysis suggests that both inter- and intrachromosomal segmental duplications have impacted on the gene count on chromosome 10. Multispecies comparative analysis indicated that we can readily annotate the protein-coding genes with current resources. We estimate that over 95% of all coding exons were identified in this study. Assessment of single base changes between the human chromosome 10 and chimpanzee sequence revealed nonsense mutations in only 21 coding genes with respect to the human sequence.

We report here the final clone map and sequence assembly of human chromosome 10. This metacentric chromosome accounts for 4.5% of the genome and is best known for harbouring the *PTEN* tumour suppressor gene and the *RET* proto-oncogene.

With the human genome sequence in hand, the task ahead is to identify the different units of genetic information embedded in the sequence and understand their function both at the molecular and cellular level. In this study we address the former by reporting a comprehensive annotation of manually inspected gene structures and their correlation to sequence variation and other features of the genomic sequence. The annotation process is assessed by comparative analysis to the genome sequence of two rodents, *Mus musculus* and *Rattus norvegicus*, and three fishes, *Tetraodon nigroviridis*, *Fugu rubripes* and *Danio rerio*. Finally, we report our preliminary findings on the distribution of single base differences along human chromosome 10 in comparison to the chimpanzee genome.

The clone map and finished sequence

A clone map spanning the euchromatic regions of the short (p) and long (q) arm of human chromosome 10 was assembled by restriction fingerprint and sequence-tagged site (STS) content analysis¹. We identified clones by screening approximately 85 genomic equivalents of P1-derived artificial chromosome (PAC), bacterial

artificial chromosome (BAC), yeast artificial chromosome (YAC), cosmid and fosmid libraries. The tiling path consists of 1,144 minimally overlapping clones (Supplementary Table S1) organized into 12 contigs (Table 1). Contig-1340 spans the entire p arm and harbours the boundary between euchromatin and heterochromatin with the proximal 250 kilobases (kb) extending into pericentromeric satellite repeats. In a detailed study of this region two further clones carrying satellite 3 sequences, contig-2069, were identified and mapped by pulse field gel electrophoresis (PFGE) ~50 kb proximal to contig-1340 (ref. 2). Similarly, contig-43 harbours the q-arm boundary with the proximal 240 kb composed of satellite repeats³. In addition, clone RP11-745D9, 94% of which consists of α -satellite repeats, was arbitrarily placed proximal to contig-43 as it was suggested to map to human chromosome 10 (E. Eichler, personal communication). A 9.75-megabase (Mb) PFGE map spanning the chromosome 10 centromere⁴ places the core α -satellite block (D10Z1) ~0.2 Mb distal to contig-102 and 1 Mb proximal to contig-43.

In contrast with the p arm, nine gaps remain in the clone map of the q arm. Five of them are clustered within a ~4-Mb region (Table 1 and Fig. 1 (rollfold); see also Supplementary Fig. S1 for a detailed view). Our inability to walk across these five gaps was due to the extensive segmental duplications in this region (Fig. 2);

however, we obtained a size estimate by fluorescent *in situ* hybridization (FISH) of RP11-172C24 (AL512595) and RP11-13E1 (AC013284) on metaphase chromosomes. We sized the remaining four gaps by FISH with clones immediately flanking each gap to extended DNA fibres. No gap was estimated to be larger than 50 kb in size. Altogether the euchromatic gaps account for no more than 840 kb (Table 1). Finally, we defined the location of both telomeres. Clone RP11-631M1 (AL713922) ends ~20 kb away from the telomeric repeats of the p arm based on the telomeric half-YAC XX-YAC2203 (<http://www.wistar.upenn.edu/Riethman/>). At the end of the q arm (qtel), clone XX-YAC2136 (BX322534) contains part of the telomeric repeat block.

Each clone of the tiling path was subjected to random subcloning and sequencing at either the Genome Therapeutics Corporation (GENE) or the Sanger Institute—the initial draft sequence of a few clones was carried out by other centres that are credited in the corresponding submissions to the EMBL/GenBank/DDBJ databases. We finished clones according to the international finishing standard (<http://genome.wustl.edu/Overview/g16stand.php>). Of the 1,144 clones in the human chromosome 10 tiling path, 221 and 913 were finished at GENE and the Sanger Institute, respectively, and three elsewhere (Supplementary Fig. S1). The remaining seven clones show persistent deletion of internal fragments. In total, we finished 131,666,441 base pairs (bp) in 18 sequence contigs; euchromatic coverage is estimated at 99.4%. Sequence accuracy was estimated as described in ref. 5 and found to exceed 99.99%. The sequence assembly comprises all known chromosome 10 messenger RNAs (RefSeq set) and STS markers from available radiation hybrid⁶ and genetic maps^{7,8} (T. Furey, personal communication). In addition, the integrity of the sequence map was independently assessed at the University of California, Santa Cruz, by alignment of fosmid and BAC paired end sequences (<http://genome.cse.ucsc.edu/>). Table 1 lists the size of each sequence contig, with the largest one spanning 44,693,577 bp.

Table 1 Clone and sequence contigs on chromosome 10

Clone contig	Sequence contig	Size (bp)	Gap size estimate (bp)
Telomere	—	—	20,000
1340	BX32259–BX276080	5,574,992	10,000
	AL365356–BX255924	12,337,510	10,000
	AL928729–AL133216	20,793,917	—
Gap	—	—	50,000
2069	AL133173–AL590623	286,100	—
Gap	—	—	50,000
Centromere	—	—	2,380,000
Gap	—	—	50,000
102	BX322613	191,752	—
Gap	—	—	50,000
43	AL031601–AC012044	3,830,268	—
Gap*	—	—	ND*
4000	AL831769–BX649215	952,201	—
Gap*	—	—	ND*
3003	AL450388–AL603965	263,307	—
Gap*	—	—	ND*
101	AL954360	163,321	—
Gap*	—	—	ND*
14	BX547991–AL731733	989,826	—
Gap*	—	—	ND*
15	AL603966–AL731572	1,941,839	10,000
	AL672187	211,435	10,000
	AL589822–AL132656	30,068,948	—
Gap	—	—	10,000
16	AC068139–AL731667	44,693,577	10,000
	AL606510–AL772134	2,696,534	—
Gap	—	—	50,000
17	BX470155–AL607044	4,615,046	—
Gap	—	—	10,000
2493 (includes telomere)	BX294094–BX511297	246,123	10,000
	AL732395–BX322534	1,809,745	—

* Cumulative gap size estimate of 750,000 for all five of the indicated gaps together. ND, not determined.

The gene and protein index

The Sanger Institute has established a standardized annotation pipeline (outlined at <http://vega.sanger.ac.uk/>) in which gene structures are drawn on the basis of human interpretation of the combined supportive evidence generated during sequence analysis. Annotation of the human chromosome 10 sequence resulted in a total of 1,787 gene structures that we then classified, as described in ref. 9, into: (1) 654 'known' genes; (2) 162 'novel genes'; (3) 219 'novel transcripts'; (4) 322 'putative genes'; and (5) 430 'pseudogenes'. Pseudogenes were further subdivided into processed (371) and unprocessed (59).

Excluding the pseudogenes, human chromosome 10 is a chromosome with an average gene density (10.4 genes Mb⁻¹). The 1,357 genes span 66,309,730 bp in total (mean 51,335 bp per gene). Therefore, 50.6% of the analysed sequence is transcribed, matching the figure reported for chromosome 22 (51%), which is gene-rich, but appearing elevated in comparison with chromosomes 6, 7, 14 and 20 (42.2%, 46.5%, 43.6% and 42.4%, respectively), which have gene densities similar to chromosome 10. The latter suggests that the human chromosome 10 genes have on average a larger genomic span than those on chromosomes 6, 7, 14 and 20. Gene size along human chromosome 10 varies enormously, with the two extremes being *CTNNA3* (1,776,209 bp) and *CALML5* (859 bp). Exons account for only 2.3% of the sequence and the mean exon size is 313 bp. The longest and shortest exons annotated in this study have a length of 9,763 (*SH3MD1*) and 3 bp (*CDH23*), respectively. *CDH23* is also the gene with most exons (69) on this chromosome. Table 2 summarizes the features of each gene class.

Alternative splicing is a major contributor to the complexity of the human transcriptome. We annotated a total of 4,204 transcripts for 1,357 gene structures (Table 2). No splice variants were annotated on the basis of alternative polyadenylation sites. Approximately 73% of the protein-coding genes have more than one transcript and 5.8 on average. For *ADD3* we annotated 22 variants. Note that the use of partial expressed sequences (for example, expressed sequence tags (ESTs)) may result in the annotation of more than one transcript per splice variant. Given this caveat, our analysis suggests a significantly higher level of alternative splicing compared with previous estimates¹⁰. Approximately 50% of the 3,456 transcripts (known and novel) do not seem to encode a protein. Annotation of these transcripts is largely based on ESTs, many of which may correspond to aberrant transcripts. Their precise role is largely unknown but they may be part of the machinery of transcriptional regulation (for example, via non-sense-mediated decay). Nevertheless, there are 1,837 transcripts with an open reading frame (ORF). Of the 342 genes with at least two transcripts having a complete ORF (that is, possess both a 5' and a 3' untranslated region (UTR)), 312 encode at least two distinct peptides.

Identification of transcription start sites and promoter regions remains a challenge in the annotation process. First, we scanned the human chromosome 10 unmasked sequence and predicted a total of

Figure 1 The sequence map of chromosome 10 and its features (see rollfold). The current sequence assembly (v1.1) and that used in the analysis presented in this study (v1.0) are available at <http://www.sanger.ac.uk/HGP/Chr10>. Tracks from top to bottom are: (1) the scale bar (in Mb); (2) the sequence map of human chromosome 10 represented by a black solid bar interrupted by clone and sequence gaps (grey); (3) syntenic blocks in the mouse (top track) and the rat (bottom track) where blocks are colour-coded per chromosome and labelled with the chromosome number and coordinates (Mb) (for example, 2: 3.1–11.3 Mb; unordered sequence contigs are tagged as random); (4) predicted CpG islands (brown); (5) regions of sequence homology to *Fugu* (blue), zebrafish (dark blue) and *Tetraodon* (dark pink); and (6) protein-coding genes. Gene names are in dark blue for the known and black for the novel CDS.

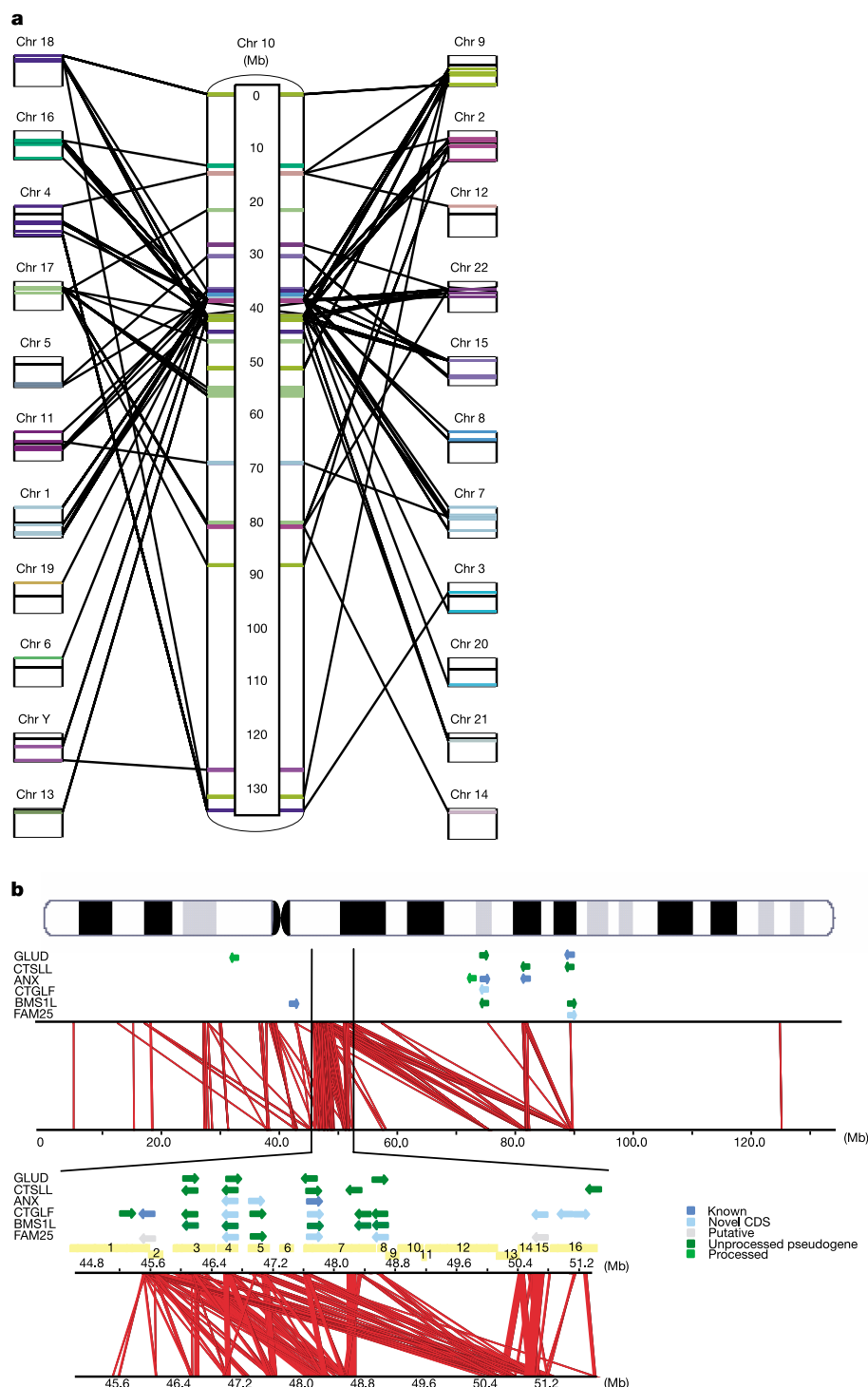


Figure 2 Chromosome 10 inter- and intrasegmental duplications. **a**, Interchromosomal duplications across chromosome 10 showing blocks of 10 kb or greater. Duplicated regions are colour-coded per chromosome and indicated as lines (thickness is proportional to physical distance). Each chromosome other than 10 is represented by an open black rectangle with a black line representing the centromere. **b**, 10q11:10q22:10q23.1:10q23.3 intrasegmental duplications. Top row, ideogram of chromosome 10; second row, positions of members of the six main gene families outside 10q11 (colour-coded per gene class; arrows indicate transcription); third row, intrachromosomal duplications across the whole chromosome showing blocks of 10 kb or greater; bottom row, enlarged view of the 10q11 region. Yellow bars represent sequence (bottom row for sequences submitted in reverse orientation) contigs between AL358394 and AL589794 (complete clone list in Supplementary Fig. S1). From left to right the

members (where parentheses indicate members only appearing in the enlarged section) of each family are: GLUD family, *GLUDP5*, (*GLUDP7*, *GLUDP8*, *GLUDP6*, *GLUDP2*), *GLUDP3*, *GLUDP1*; CTSL family, (*CTSL5*, *CTSL7*, *CTSL2*, *CTSL3*, *CTSL4*), *CTSL6*, *CTSL1*; ANXA family, (*ANXA8L1*, *ANXA8L2*, *ANXA8*), *ANXA2P3*, *ANXA7*, *ANXA11*; CTGLF family, (*CTGLF10P*, *CTGLF1*, *CTGLF13P*, *CTGLF7*, *CTGLF11P*, *CTGLF6*, *CTGLF9P*, *CTGLF12P*, *CTGLF5*, *CTGLF4*, *CTGLF3*), *CTGLF2*; BMS1L family, *BMS1L*, (*BMSILP1*, *BMSILP2*, *BMSILP6*, *BMSILP5*, *BMSILP7*), *BMSILP4*, *BMSILP3*; FAM25 family, (*FAM25E*, *FAM25B*, *FAM25HP*, *FAM25G*, *FAM25C*, *FAM25D*), *FAM25A*. *ANXA7* and *ANXA11* were included owing to their proximity to the 10q22 and 10q23.1 locus, respectively. Seven *CTGLF1* paralogues (Supplementary Table S3) were annotated as novel genes for consistency but probably represent expressed pseudogenes.

Table 2 **Structural characteristics of annotated gene structures on chromosome 10**

Gene structure class	Number of genes	Total transcribed length (bp)	Mean gene length (bp)	Mean exon length (bp)	Mean transcripts per gene	Mean exons per gene
Known genes	654	56,312,218	86,870	316	4.76	11.02
Novel CDS	162	3,615,155	22,435	344	2.12	4.96
Total protein coding	816	59,482,121	74,078	320	4.24	9.81
Novel transcript	219	6,123,588	29,166	272	1.77	4.06
Putative genes	322	2,798,239	8,780	287	1.12	2.15
Processed pseudogenes	371	328,131	887	638	1.00	1.19
Unprocessed	59	982,662	16,655	176	1.00	6.95
Total structures	1,787	67,353,917	39,717	322	2.59	5.84

Note that transcribed length is not additive because some genes overlap. CDS, coding sequence.

1,025 CpG islands (Fig. 1), which are known to be associated with the 5' end of an estimated 60% of human genes¹¹. We then used Eponine¹² to predict transcription start sites and FirstEF¹³ to predict regions that encompass the promoter and 5' exon. In total, FirstEF and Eponine predicted 1,801 (rank = 1, score ≥ 0.8) and 2,800 features, respectively (Supplementary Fig. S1). Notably, 62% of FirstEF and 96% of Eponine features directly overlapped CpG islands, suggesting a heavy bias towards this feature in both algorithms. The distribution of CpG islands, FirstEF and Eponine hits was also assessed relative to the first exon of each of the 4,635 annotated transcripts using a window of 5,000 bp upstream and 1,000 bp downstream of the exon. Table 3 summarizes the results obtained per gene class. For example, in the 'known' class, 1,544 (49.6%) and 1,124 (36.1%) transcripts had a FirstEF and Eponine feature, respectively. Not surprisingly, 89.9% of FirstEF and 96.7% of Eponine transcripts were also associated with a CpG island. Note that Eponine predicts multiple transcription start sites per transcript (4.24 on average).

Regulation of gene expression by antisense transcription is a recognized mechanism with examples reported in species ranging from bacteria to mammals^{14–16}. We observed widespread occurrence of overlapping coding genes (either strand) in human chromosome 10 (101 pairs in total). In 38 cases one of the genes is fully contained within an intron of the other, typically on the opposite strand (34). For example, the second intron of the splice variant LIPA-004 encompasses four members of the IFIT gene cluster (Supplementary Table S3 and Fig. S2) and appears to be transcribed from the same bidirectional promoter as *IFIT5* (IFIT cluster member). Interestingly, the LIPA-001 and -002 variants do not overlap any IFIT gene, whereas LIPA-003 and LIPA-005 both overlap with *IFIT2* and *IFIT4* (Supplementary Fig. S2). Mutations in *LIPA* can cause Wolman and cholesteryl ester storage disease (Online Mendelian Inheritance in Man (OMIM) 278000). Among partially overlapping pairs (opposite strands), 34% involve the respective 5' exons, which is indicative in each case of a bidirectional promoter. We also searched for non-coding transcripts located on the opposite strand of coding genes. There are 67 antisense transcripts overlapping 63 coding genes. The two most common patterns observed were intragenic with partial overlap of one exon, and partial overlap with the most 5' exon of the coding gene. For *ZNF32*, we found two antisense transcripts (*ZNF32OS1* and *ZNF32OS2*).

Finally, we looked at the distribution of known protein domains

in both human chromosome 10 (this study) and the whole genome (Ensembl v.17.33.1) proteome using InterProScan. At the gene level, 70.6% of the human chromosome 10 peptides have at least one InterPro match with a Pfam domain and 32% are multidomain (1.37 distinct InterPros on average). Supplementary Table S2 shows the top 24 domains in chromosome 10 alongside their genome-wide ranking, suggesting that this chromosome is enriched in peptides with a lipase (IPR000734), aldo/keto reductase (IPR001395) or alpha/beta hydrolase (IPR000073) domain. BLASTP analysis (e -values $< 10^{-15}$) showed that all six genes encoding the peptides carrying the IPR001395 domain are clustered (AKRIC cluster) at position 4.8–5.3 Mb, whereas there are two lipase clusters, LIP (90.0–91.0 Mb; six members) and PNLIP (117.9–118.1 Mb; four members). In total, we found 42 gene clusters along human chromosome 10 (Supplementary Table S3).

Genomic landscape

The average G+C and repeat content of human chromosome 10 are 41.58% and 43.66%, respectively. The distribution of the main classes of repeats (Supplementary Table S4), the G+C and CpG density plots all seem to follow the known genome-wide trends. For example, the G+C content fluctuates along the chromosome, peaks at the qtl and is positively correlated to gene density (Supplementary Fig. S1). Large genes tend to be located adjacent to or within gene-poor regions, for example *PRKG1* and *PCDH15* (interval 52.0–59.0 Mb), whereas regions of high gene density seem enriched in short interspersed elements (SINES).

Segmental duplications are an important feature of the genomic landscape, being an integral part of the evolutionary machinery. Figure 2a illustrates the interchromosomal duplications along human chromosome 10 (see also ref. 17 and <http://humanparalogy.cwru.edu/SDD>). Figure 2b shows the extensive segmental duplications within a 5 Mb region at 10q11 with sequences further dispersed at 10q22, 10q23.1 and 10q23.2. Using the draft genome sequence Crosier and colleagues¹⁸ reported that 10q11 has been subject to multiple rounds of local duplications in at least the last 30–40 million years (Myr); they also showed that a 10q11:10q23 paracentric inversion occurred after the divergence of orang-utan from other great apes and hypothesized that the 10q22 locus resulted from chromosome-specific duplicative transposition. Bryce and colleagues¹⁹ characterized three cathepsin-L-like paralogues, which are expressed pseudogenes, and reported FISH signals

Table 3 **Correlation of CpG island, FirstEF and Eponine features with annotated transcripts**

Gene class	All Transcripts		FirstEF transcripts		Eponine transcripts	
	Complete	Incomplete	Complete	Incomplete	Complete	Incomplete
Known	1,490 (580)	1,623 (1,018)	593 (531)	951 (877)	421 (412)	703 (682)
Novel CDS	130 (57)	213 (75)	45 (43)	62 (57)	38 (38)	43 (41)
Novel transcript	388 (119)	0 (0)	120 (105)	0 (0)	83 (79)	0 (0)
Processed pseudogene	371 (32)	0 (0)	29 (22)	0 (0)	21 (19)	0 (0)
Unprocessed pseudogene	59 (11)	1 (2)	14 (10)	1 (1)	9 (8)	0 (0)
Putative	360 (74)	0 (0)	87 (64)	0 (0)	57 (50)	0 (0)

Values in parentheses are for transcripts associated with a CpG island.

at 10q11, 10q22 and 10q23, a pattern previously seen with the *GLUD* paralogues²⁰ (Fig. 2b). The duplication of the *CTSL* locus between chromosomes 9 and 10 was estimated to have occurred some 40 Myr ago¹⁹. We identified four and three additional *CTSL* and *GLUD* pseudogenes, respectively (Fig. 2b), consistent with the local duplication events involving *BMS1L* (known as *KIAA0187*)¹⁸ and *CTGLF1* (this study). In total, we annotated 7 *BMS1L* and 11 *CTGLF1* paralogues (Fig. 2b; functional genes in dark blue). Retro-position of a truncated *KIAA1099* (*CENTG2*; chromosome 2) mRNA gave rise to a processed pseudogene on chromosome 10 (ref. 18). However, this pseudogene forms the 3' exon of *CTGLF1*, suggesting that this gene resulted from a fusion between an ancestral gene and the *CENTG2* pseudogene, and the retroposition event predated all segmental duplications. Note that there is also evidence of transcripts combining exons of *CTGLF1* and *BMS1L* paralogues. Interestingly, we predicted a novel gene, *FAM25A* (based on mouse complementary DNA AK008614 and without similarity to any known protein), with seven human chromosome 10 paralogues (Fig. 2b) that follow the pattern of *GLUD*, *BMS1L*, *CTGLF* and *CTSL*. In addition to these five types of low copy repeats, the segmental duplications in 10q11:10q22:10q23 seem to have impacted on the number of genes on this chromosome. Notably, 31% of all the functionally related gene clusters are located within 10q11 and 10q23 (Supplementary Fig. S1 and Table S3).

The average recombination rate across the chromosome is 1.32 cM Mb⁻¹ (Fig. 3, sex average). Note that ref. 7 used the draft human chromosome 10 sequence (inflated by ~8%) and thus obtained a lower figure (1.21 cM Mb⁻¹). The rate of male recombination is higher than the female rate near the telomeres, whereas between D10S211 and D10S575 the female rate is higher (Fig. 3). This comparison also indicates the presence of two female-specific recombination hotspots (Fig. 3, arrows). As expected, pericentromeric regions display a low rate of recombination, more than twofold below the chromosome average. In particular, the region between D10S1247 and D10S1783 has a rate of 0.3 cM Mb⁻¹ and contains the only human chromosome 10 recombination 'desert'²¹. The region of extensive segmental duplications at 10q11 shows a low

rate of recombination (0.72 cM Mb⁻¹). Thus, meiotic recombination is unlikely to have been the driving force in generating these duplications. Our analysis confirmed the recombination 'jungle' between D10S1782–D10S1651, which we extended to D10S212 (3.4 cM Mb⁻¹); we refuted the one between D10S189 and D10S1728 (2.3 cM Mb⁻¹), and identified a new one between D10S1154 and D10S552 (4.55 cM Mb⁻¹).

Comparative sequence analysis

Many of the functional units in present-day vertebrate genomes have been conserved through evolution. Coding regions are highly conserved across all vertebrates, whereas non-coding regions are conserved among more closely related species. The genome sequences of three fishes (*Tetraodon*, *Fugu* and zebrafish) and two rodents (mouse and rat) were publicly available at the time of analysis. We compared the sequence of human chromosome 10 to each of the above species and searched for conserved regions with coding potential—distinguishing functional non-coding elements above background sequence conservation requires the use of additional genomes²². We then correlated the obtained hits in each species with the annotated genes. As expected, each of the fish genomes showed higher specificity (>0.82; that is, 82% of sequences conserved in fish overlapped human annotated exons) than the rodents (>0.29), whereas the highest specificity was obtained using all genomes together (0.96). Sensitivity was higher using a rodent genome (0.7; that is, 70% of all annotated exons had matches in rodent) than a fish genome (0.4).

Of the 1,787 gene structures (16,765 exons), 84% (78% of exons) had at least one exon supported by a conserved region in one of the other genomes and 52% (32% of exons) in all genomes. Note that the figures given for exons are underestimated owing to lower sequence conservation in the untranslated regions. Protein-coding genes are highly conserved (98.5%; 85% of exons). In contrast, 61% (29% of exons) and 45% (27% of exons) of novel transcripts and putative genes, respectively, have at least one match. Furthermore, only 21% of novel transcripts and 8% of putative genes show conservation with a fish (91% for protein-coding genes). Typically, novel transcripts were annotated on the basis of solid experimental evidence (that is, human mRNA) and may represent either genes that have evolved more rapidly or non-coding RNAs.

On the basis of specificity, regions conserved in all six species can serve as a measure of completeness of the gene annotation process that occurred independently of the comparative analysis. We found 5,604 such evolutionarily conserved regions (ECRs) of which 5,292 mapped inside annotated exons (including pseudogenes). Of the remaining 312 ECRs, 142 were intergenic and 170 intragenic. On inspection, we found 79 of these ECRs with supportive evidence to annotate a missed exon, most of which were part of a pseudogene (79%). The remaining 233 ECRs provide the basis to estimate that we have annotated at least 95.8% of all conserved coding exons on human chromosome 10. This is a conservative estimate as 131 of these ECRs are located in introns and may represent conserved non-exonic sequences. Interestingly, 54 (41%) of them are associated with just four genes: *C10orf11* (26; also known as *CDA017*), *EBF3* (20), *TCF7L2* (5) and *PAX2* (3). All but *C10orf11* (unknown function) are transcription factors. Figure 4 shows a MultiPip-Maker²³ alignment of the orthologous *EBF3* loci and the relative position of ECRs in the human gene. Note that sequence identity is often higher in ECRs than in exons.

Sequence variation

During the human chromosome 10 project we discovered 35,882 single nucleotide polymorphisms (SNPs) by sequence alignment in regions of clone overlaps. In total, we mapped 143,364 SNPs (dbSNP release 115) to the chromosome 10 sequence. Supplementary Fig. S1 shows the density plots for randomly discovered²⁴ and all SNPs across the chromosome.

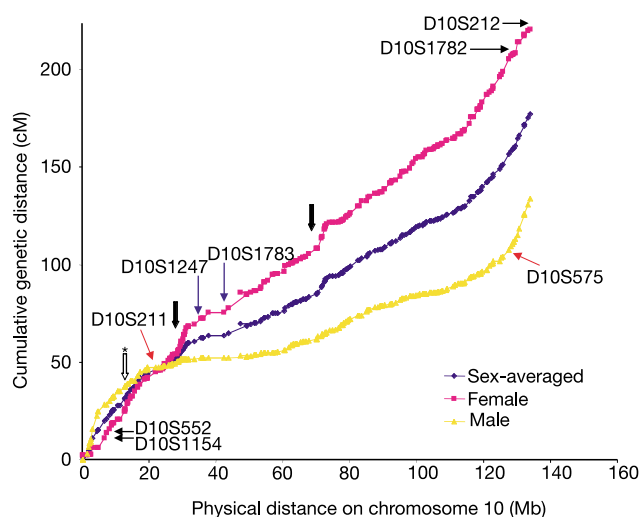


Figure 3 Alignment of the deCODE genetic map of chromosome 10 to the physical map from the telomeric end of the short arm to the telomeric end of the long arm. The position of each genetic marker on the female, male and sex-averaged genetic map is indicated. Female-specific recombination hotspots are indicated by thick arrows (left, D10S1732–D10S208; right, D10S599–D10S676). The location of markers flanking recombination deserts (blue arrows) and jungles (black arrows) is shown. The asterisk indicates the location of the refuted jungle (D10S189–D10S1728).

There are 5,864 (4.1%) exonic and 65,973 (46%) intronic SNPs. Of the 1,821 SNPs in coding exons 984 are non-synonymous. *MSMB* has the most polymorphic coding region with 43 SNPs kb⁻¹; it encodes a protein with inhibin-like activity and its expression is decreased in prostate cancer²⁵.

We also considered 729,553 human–chimpanzee single base differences (SBDs) remapped on the current assembly of human chromosome 10. These were high-confidence sequence differences originally identified by aligning 14 million shotgun reads of the chimpanzee genome, generated jointly by the Whitehead Institute and Washington University Genome Centers, to the human genome sequence assembly (NCBI build 31). We first removed all human–chimpanzee SBDs that co-localized with known human SNPs. Supplementary Fig. S1 shows the density plot of the remaining 703,338 SBDs. Of those, 55.3% are intergenic, 42.9% intronic and 1.8% exonic. The highest density of human–chimpanzee SBDs, fourfold greater than the average level, was observed in a 200-kb gene-poor region at 19.43–19.63 Mb. We then examined the 12,710 human–chimpanzee SBDs that lie in exons of the 816 human coding genes. Of those, 3,972 were in coding regions and can be subdivided further into 2,273 synonymous, 1,678 non-synonymous and 21 nonsense with respect to the human sequence. For each gene we calculated the rate of evolution of non-synonymous (K_a) and synonymous (K_s) changes, and the ratio K_a/K_s , which provides a measure of evolutionary selection. Supplementary Table S5 lists the 1,413 transcripts with at least one coding human–chimpanzee SBD sorted on the K_a/K_s value. There are only 29 transcripts (21 genes) that have a K_a/K_s value ≥ 1 , whereas there are 484 without non-synonymous SBDs. Note that several caveats apply in this type of analysis owing to the incomplete nature of both the chimpanzee data and the list of human SNPs; we used the number of intronic human–chimpanzee SBDs per base in comparison to the chromosome average of 0.005 as a possible estimate of coverage. The gene with most non-synonymous human–chimpanzee SBDs is *MKI67*, an antigen identified by monoclonal antibody Ki-67, which appears to be fast evolving in humans ($K_a/K_s = 1.038507$; SNP data). The expression pattern of *MKI67* in gastric and other cancers is under investigation as this gene is expressed in proliferating cells. Interestingly, a nonsense human–chimpanzee SBD is present in both of its coding transcripts. Among the 21 genes with nonsense human–chimpanzee SBDs notable examples are the serotonin receptor *HTR7* (the neurotransmitter serotonin is thought to be involved

in cognition and behaviour), *PSAP* (prosaposin; involved in variant Gaucher's disease and metachromatic leukodystrophy) and the developmental gene *NODAL*.

Medical implications

At the time of writing there were 85 disease loci reported on human chromosome 10 (<http://www.ncbi.nlm.nih.gov/omim/>), a 47% increase since 1999 (ref. 26). Several of these loci account for multiple disease phenotypes caused by mutations in the same gene; notable examples are *FGFR2* (OMIM 176943), *PTEN* (OMIM 601728) and the proto-oncogene *RET* (OMIM 164761). Since *PTEN* was first shown to be mutated in brain, breast and prostate cancers²⁷, there has been an explosion of reported mutations (110 germline and 332 somatic mutations)²⁸ and disease phenotypes²⁹. Human chromosome 10 harbours several other genes involved in tumorigenesis; for example, deregulation of *TLX1*, *NFKB2* or *BM11* caused by chromosomal translocations or amplifications has been detected in lymphatic neoplasms. Mapping of allelic imbalances and functional studies suggest the presence of additional tumour-related genes. The finished and annotated sequence is key in the process of cloning these and other hitherto unidentified disease-associated genes.

The prompt release of both the clone and sequence map resources throughout the project has accelerated the cloning of many disease-causing genes. To this end we recently showed as part of the European ADLITE consortium that mutations in the *LGII* gene cause autosomal dominant lateral temporal epilepsy³⁰. Notably, we found that the *FRA10A* folate-sensitive fragile site is located close to *LGII* and its expression is associated with the expansion of a polymorphic CGG repeat located at the 5' UTR of *FRA10AC1*, a gene encoding a novel nuclear protein³¹. There are seven fragile sites mapping to human chromosome 10 (ref. 32).

The challenge ahead is to unravel the molecular basis of common disease. An increasing number of susceptibility loci for complex diseases is being mapped to human chromosome 10, including metabolic diseases such as type I diabetes (IDDM10 and IDDM17), psychiatric disorders such as schizophrenia, or neurodegenerative illnesses such as Alzheimer's disease^{26,32}. In a case control study of morbidly obese and healthy individuals Boutin and colleagues³³ identified a SNP in the *GAD2* gene that increases the risk for obesity as well as a protective haplotype. Studies so far have mainly focused on candidate genes. The construction of a comprehensive haplotype

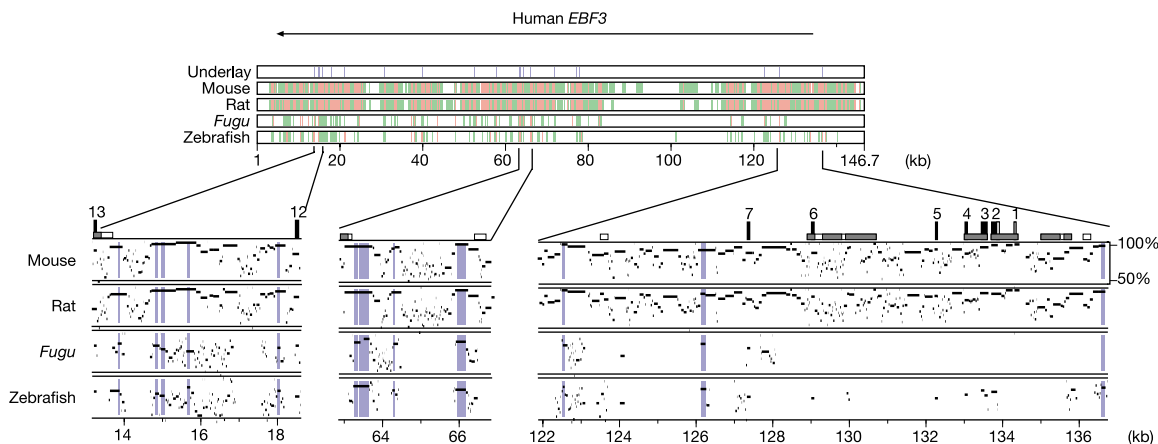


Figure 4 Multispecies alignment of orthologous *EBF3* loci. The human early B-cell factor 3 (*EBF3*) gene is represented by the arrow at the top. Alignments are displayed using MultiPipMaker²³. In the top panel, the first track shows the location of the ECRs (blue lines) across the human locus, whereas the following four tracks show regions conserved in mouse, rat, *Fugu* and zebrafish, respectively (green, aligned regions; orange, aligned

regions with at least 70% nucleotide identity and no gap over 100 bp). The bottom panel shows a detailed view of the three regions with the highest number of ECRs. Vertical black and grey numbered boxes represent coding and UTR exons, respectively. The scale at the right indicates the percentage of sequence identity. Physical distance is given in kilobases (kb).

map of the human genome is well underway³⁴, making it possible to undertake a systematic approach to scanning the genome for associations to disease-related and other phenotypes. □

Methods

The mapping and sequencing methods used in the assembly of the bacterial clone and sequence map of chromosome 10, respectively, as well as the tools of the gene annotation pipeline are described in refs 1 and 35 (see ref. 36 for detailed protocols). Manual annotation of gene structures followed the guidelines agreed in the human annotation workshop (HAWK; <http://www.sanger.ac.uk/HGP/havana/hawk.shtml>), whereas gene symbols were approved where possible by the HUGO Gene Nomenclature Committee³⁷. Protein translations were analysed with InterProScan (<http://www.ebi.ac.uk/interpro/scan.html>), which was run via the Ensembl protein annotation pipeline, to obtain Pfam, Prosite, Prints and Profiles domain matches.

Alignments for inter- and intrachromosomal duplications were performed with WU-BLASTN (<http://blast.wustl.edu>) using the current sequence assembly of chromosome 10 and the NCBI34 build for the rest of the genome. All sequences were repeat masked with RepeatMasker (<http://repeatmasker.genome.washington.edu>) and low-quality alignments (e -value $>10^{-30}$, sequence identity $<90\%$, length <80 bp) were discarded. For intrachromosomal duplications, self matches were discarded. For interchromosomal duplications, the sequence was split into 400-kb segments. Adjacent matches in the same orientation were joined together as described by ref. 38. Only blocks of 10 kb or greater were retained.

The following sequence assemblies were used for comparative analysis: *M. musculus* NCBI build 30 (Mouse Genome Sequencing Consortium; http://www.ensembl.org/Mus_musculus/resources.html); *R. norvegicus* version 2.0 (Rat Genome Sequencing Consortium; <http://www.hgsc.bcm.tmc.edu/projects/rat/>); *D. rerio* Assembly version 1 (Sanger Institute; http://www.sanger.ac.uk/Projects/D_rerio/); *F. rubripes* version 2 (International Fugu Genome Consortium; <http://www.fugu-sg.org/project/info.html>); and *T. nigroviridis* version 6 (Genoscope and Whitehead Institute for Genome Research; <http://www.genoscope.cns.fr/externe/tetraodon/Ressource.html>). The repeat-masked sequence of chromosome 10 was aligned to the mouse and rat genome sequences using BLASTZ³⁹ and the resulting matches were post-processed by axtBest and subsetAxt (W. J. Kent; <http://www.soe.ucsc.edu/~kent/src>) as described elsewhere³⁵. Alignment of the three fish genome sequences to chromosome 10 was performed with WU-TBLASTX using the scoring matrix, parameters and filtering strategy applied by Exofish⁴⁰. Overlapping alignments to different sequences were merged to produce contiguous regions of sequence conservation, analogous to the ECRs or 'Ecores', reported by Exofish.

SNPs in sequence overlaps were identified using a modification of the SSAHA software⁴¹. The chromosome 10 SNPs (dbSNP release 115) were mapped to the sequence assembly of this chromosome (this study) first with SSAHA and then Cross-Match. Of the approximately 14 million chimpanzee reads (<http://www.genome.gov/11008056>) mapped onto the human sequence assembly (NCBI build 31), those mapping to chromosome 10 were remapped to our sequence assembly and used to identify human-chimpanzee SBDs.

Received 17 December 2003; accepted 9 March 2004; doi:10.1038/nature02462.

1. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409**, 942–943 (2001).
2. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13**, 159–172 (2003).
3. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9**, 2029–2042 (2000).
4. Jackson, M. S., See, C. G., Mulligan, L. M. & Lauffart, B. F. A 9.75-Mb map across the centromere of human chromosome 10. *Genomics* **33**, 258–270 (1996).
5. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9**, 1–4 (1999).
6. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
7. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
8. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
9. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
10. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
12. Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
13. Davuluri, R. V., Grosse, I. & Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nature Genet.* **29**, 412–417 (2001).

14. Wagner, E. G. & Simons, R. W. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48**, 713–742 (1994).
15. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929 (2001).
16. Kiyosawa, H. *et al.* Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13**, 1324–1334 (2003).
17. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
18. Crosier, M. *et al.* Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. *Genome Res.* **12**, 67–80 (2002).
19. Bryce, S. D. *et al.* A novel family of cathepsin L-like (CTSLL) sequences on human chromosome 10q and related transcripts. *Genomics* **24**, 568–576 (1994).
20. Deloukas, P., Dauwerse, J. G., Moschonas, N. K., van Ommen, G. J. & van Loon, A. P. Three human glutamate dehydrogenase genes (GLUD1, GLUDP2, and GLUDP3) are located on chromosome 10q, but are not closely physically linked. *Genomics* **17**, 676–681 (1993).
21. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
22. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
23. Schwartz, S. *et al.* PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).
24. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
25. Vanaja, D. K., Chevillet, J. C., Iturria, S. J. & Young, C. Y. Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. *Cancer Res.* **63**, 3877–3882 (2003).
26. Deloukas, P., French, L., Meitinger, T. & Moschonas, N. K. Report of the third international workshop on human chromosome 10 mapping and sequencing 1999. *Cytogenet. Cell Genet.* **90**, 1–12 (2000).
27. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1946 (1997).
28. Bonneau, D. & Longy, M. Mutations of the human PTEN gene. *Hum. Mutat.* **16**, 109–122 (2000).
29. Eng, C. PTEN: one gene, many syndromes. *Hum. Mutat.* **22**, 183–198 (2003).
30. Morante-Redolat, J. M. *et al.* Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy. *Hum. Mol. Genet.* **11**, 1119–1128 (2002).
31. Sarafidou, T. *et al.* Folate-sensitive fragile site FRA10A is due to an expansion of a CGG-repeat in a novel gene FRA10AC1, encoding a nuclear protein. *Genomics* (in the press).
32. Moschonas, N. K. in *Encyclopedia of the Human Genome* Vol. 1, 618–625 (Nature Publishing Group, London, 2003).
33. Boutin, P. *et al.* GAD2 on chromosome 10p12 is a candidate gene for human obesity. *PLoS Biol.* **1**, 1–11 (2003).
34. The International HapMap Consortium. The International HapMap project. *Nature* **426**, 789–796 (2003).
35. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
36. Dunham, I. (ed.) *Genome Mapping and Sequencing* (Horizon Scientific, Wymondham, UK).
37. Wain, H. M., Lush, M., Ducluzeau, F. & Povey, S. Genew: the human gene nomenclature database. *Nucleic Acids Res.* **30**, 169–171 (2002).
38. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
39. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
40. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
41. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Fugu, Mouse and Rat Genome Sequencing Consortia; Genoscope and Whitehead Institute for Genome Research (Tetraodon genome data); the members of the zebrafish project at the Sanger Institute; the groups at the Washington University Genome Sequencing Center and the Whitehead Institute for early access to the chimpanzee whole-genome shotgun data; the laboratories of H. Riethman and M. S. Jackson for the provision of telomeric and pericentromeric clones, respectively; H. M. Wain, E. A. Bruford, C. C. Talbot, V. K. Khodiyar, M. J. Lush, M. W. Wright and S. Povey for assigning official gene symbols; the EMBL, GenBank and Ensembl database groups; I. Dunham for critical reading of the manuscript; and the Wellcome Trust for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.D. (panos@sanger.ac.uk). All DNA sequences reported in this study have been deposited in the EMBL or GenBank databases and the accession numbers can be found in Supplementary Fig. S1.

DNA sequence and comparative analysis of chimpanzee chromosome 22

The International Chimpanzee Chromosome 22 Consortium*

*A list of authors and their affiliations appears at the end of the paper

Human–chimpanzee comparative genome research is essential for narrowing down genetic changes involved in the acquisition of unique human features, such as highly developed cognitive functions, bipedalism or the use of complex language. Here, we report the high-quality DNA sequence of 33.3 megabases of chimpanzee chromosome 22. By comparing the whole sequence with the human counterpart, chromosome 21, we found that 1.44% of the chromosome consists of single-base substitutions in addition to nearly 68,000 insertions or deletions. These differences are sufficient to generate changes in most of the proteins. Indeed, 83% of the 231 coding sequences, including functionally important genes, show differences at the amino acid sequence level. Furthermore, we demonstrate different expansion of particular subfamilies of retrotransposons between the lineages, suggesting different impacts of retrotranspositions on human and chimpanzee evolution. The genomic changes after speciation and their biological consequences seem more complex than originally hypothesized.

To understand the genetic basis of the unique features of humans, a number of pilot studies comparing the human and chimpanzee genomes have been conducted^{1–5}. Estimates of nucleotide substitution rates of aligned sequences range from 1.23% by bacterial artificial chromosome (BAC) end sequencing³ to about 2% by molecular analysis^{1,6–8}, whereas the overall sequence difference was estimated to be approximately 5% by taking regions of insertions or deletions (indels) into account⁹. Chromosomal rearrangements including duplications, translocations and transpositions have also been identified^{10,11}. However, owing to technological limitations there is not an integrated picture of the dynamic changes of the genome, thus a gold standard is required to evaluate the overall consequence of these genetic changes on human evolution.

To address these issues and to be able to detect molecular blueprints that have shaped the two genomes, we have conducted a human–chimpanzee whole-chromosome comparison at the nucleotide sequence level on human chromosome 21 (HSA21) and its orthologue chimpanzee chromosome 22 (PTR22). HSA21 is one of the most well characterized human chromosomes^{7,12–14} and serves as a representative of the human genome by having characteristic features such as uneven distribution of G+C content with a high correlation to gene density, and repetitive/duplicated structures, allowing for detailed long-range comparative studies with PTR22. Moreover, molecular analysis of HSA21 and its genes is of central medical interest because of trisomy 21, the most common genetic cause of mental retardation in the human population. One case of trisomy 22 in chimpanzee has been reported, with phenotypic features similar to human Down's syndrome¹⁵. Therefore, our analysis of these chromosomes should reveal dynamic changes that may reflect general evolutionary events occurring throughout the human genome.

Mapping, sequencing and overview of PTR22

We used three different BAC libraries prepared from genomic DNA originating from three male chimpanzees (*Pan troglodytes*). Sequence coverage of the euchromatic portion of the long arm of chromosome 22 (PTR22q) is estimated to be 98.6% (33.3 megabases (Mb)). Accuracy was calculated as 99.9983% from the overlapping clone sequences and $\geq 99.9981\%$ on the basis of Phrap scores¹⁶. Altogether, these efforts enabled us to produce a sequence with the highest possible accuracy to be used for reliable comparative analysis (see Supplementary Information for details).

The overall structural features of PTR22q are almost the same as those of HSA21q. The G+C content of these chromosomes is around 41% (Table 1). The corresponding regions between HSA21q and PTR22q, where the extra regions (see Methods) are excluded, show a roughly 400-kilobase (kb) or 1.2% difference in size, with HSA21q being larger than PTR22q. The difference is mainly due to interspersed repeats (ISRs) and simple repeats, representing 63.2% (53.7% and 9.5%, respectively) of the regions corresponding to the gaps in PTR22q. The pericentromeric copy of a 200-kb region found duplicated in HSA21q is missing in PTR22q, as reported previously¹⁷. We also detected human-specific sequences that are neither repetitive nor low complexity and are unique in the nr data set of NCBI (<http://www.ncbi.nih.gov/>). For example, a 1,245-base-pair (bp) insertion found in the first intron of

Table 1 Statistics on HSA21q and PTR22q

Genome characteristic	HSA21q	PTR22q		
Size (bp)*	33,127,944	32,799,845		
Unaligned sites†	25,242	101,709		
Sequencing gaps	14	22		
Clone gaps‡	3	2		
Estimated total clone gap size	73,108	74,311		
G+C content (%)	40.94	41.01		
CG dinucleotides	361,259	358,450		
CpG islands	950	885		
Nucleotide diversity (%)	0.072	0.14		
Repeats	HSA21q		PTR22q	
	Bp	Number	Bp	Number
SINEs	3,649,153	15,137	3,614,825	15,048
Young Alu elements§	21,557	75	2,606	10
LINEs	5,853,821	8,737	5,736,911	8,673
Young L1 elements	82,493	48	78,657	55
LTRs	3,621,501	7,282	3,550,807	7,180
Transposons	949,215	3,363	945,129	3,350
RNAs¶	8,830	100	8,722	99
Satellites	19,327	21	14,773	18
Others	30,452	38	34,776	43
Total	14,132,299		13,905,943	
	42.7%		42.4%	

*Size of the contig data after the site where the first base of the PTR22q contig is aligned.

†Regions extended into HSA21q clone gaps and subtelomeric unmatched regions.

‡Excluding pericentromeric and subtelomeric gaps.

§AluYa5, AluYa8, AluYb8 and AluYb9.

||L1Hs and L1PA2.

¶Small nuclear RNA, small cytoplasmic RNA, 5S ribosomal RNA, transfer RNA, 7SL RNA and other small RNA genes.

PFKL in HSA21q was confirmed to be human-specific and even locus-specific by searching for similar sequences in the nr database. Four expressed sequence tag (EST)/complementary DNA (cDNA) sequences (BQ711940, BQ706616 and AA453553 on the *PFKL* strand and AJ003358 on the complementary strand) are mapped specifically onto this region.

We did not observe significant correlations between the frequencies of base substitution and indels. Two large indel hotspots were found at around 9.5–11.5 Mb and 16.5–17.5 Mb from the centromere, as previously suggested^{3,14} (Fig. 1b). In addition, we found large human insertions/chimpanzee deletions in the first introns of the *NCAM2* (~10 kb) and *GRIK1* (~4 kb) genes, which are both related to neural functions.

One of the largest structural changes identified is a 54-kb region located 11.4 Mb from the centromere in HSA21q but that is absent in PTR22q. This region is flanked by HSAT5 satellite repeats and consists of 164 fragments from 64 different long terminal repeat (LTR) elements with inversions of small portions at even intervals, suggesting rearrangement driven by interspersed repetitive elements. The most distal 25,242-bp region of HSA21q ending with a TTAGGG telomeric repeat did not align to the last 80,879 bp of PTR22q, which instead shares high similarity to human chromosomes 2, 9 and 10. This suggests that the subtelomeric region of PTR22q might be larger than that of HSA21q. The pericentromeric regions of these chromosomes have complex structures: those of PTR22q show a high degree of similarity to human acrocentric chromosomes 13, 15, and 18 (submetacentric only in humans but acrocentric in all great apes) as well as HSA21 (data not shown).

Base substitutions

The overall nucleotide substitution level in aligned regions between PTR22q and HSA21q is about 1.44% (excluding indels), which is significantly higher than the calculation based on high-quality BAC end sequence data (1.23%)³. Similar trends have also been observed for HSA21q based on smaller comparisons with the chimpanzee genome¹⁸.

The distribution of base substitution rate along the chromosome

is shown in Fig. 1a. The frequency of base substitution is distributed around 1.44% along the chromosome, except for elevated regions in the pericentromeric and subtelomeric regions. The most conserved region was at about the 12.5-Mb region (0.87%, over 100 kb), corresponding to the distal boundary region of the gene desert¹². Notably, no protein-coding genes have been identified in this highly conserved region. The correlation between the G+C content and the base substitution rate increases along the chromosome, and is especially high in the last 5 Mb of the telomeric region of PTR22q (data not shown). The frequency of base substitution in repetitive sequences also tends to vary, increasing from CR1/long interspersed element (LINE) (divergence, 1.13%; G+C percentage, 39.0%) and L1/LINE repeats (1.38%; 35.6%) to Alu/short interspersed element (SINE) (1.81%; 51.7%), ERVK/LTR (1.88%; 44.6%), CpG islands (2.26%; 65.1%) and simple repeats (4.06%; 44.8%), as discussed previously¹⁹. The CG dinucleotide frequencies are significantly ($P < 0.01$) different between PTR22q and HSA21q (Table 1).

Repetitive elements

As described above, HSA21q is about 1.2% longer in size than PTR22q. This difference can mostly be explained by the fact that several subfamilies of transposable elements, such as L1Hs (11 versus 2), MER83B (11 versus 0), AluYa5 (23 versus 3) and AluYb8 (37 versus 2), are more common in human than in chimpanzee^{20–23}. Five LTR subfamilies (LTR/ERV1) are more abundant in HSA21q. All MER4A1-int and MER83B-int elements are specific to HSA21q and are clustered in limited regions (positions 21,173,756–21,180,021 bp and 40,651,714–40,657,562 bp, respectively) with overlapping tandem repeats.

According to the sequence alignments, single copies of L1Hs and AluYa5 and two copies of AluYb8 are found in both HSA21q and PTR22q, indicating their presence in the genome of the last common ancestor. All of the seven AluYb9 elements found in HSA21q and the one in PTR22q are lineage-specific, suggesting that these elements have been integrated after speciation. Although the AluYa8 subfamily is thought to be a recent derivative of AluYa5 (ref. 24), we found a few AluYa8 units in both species.

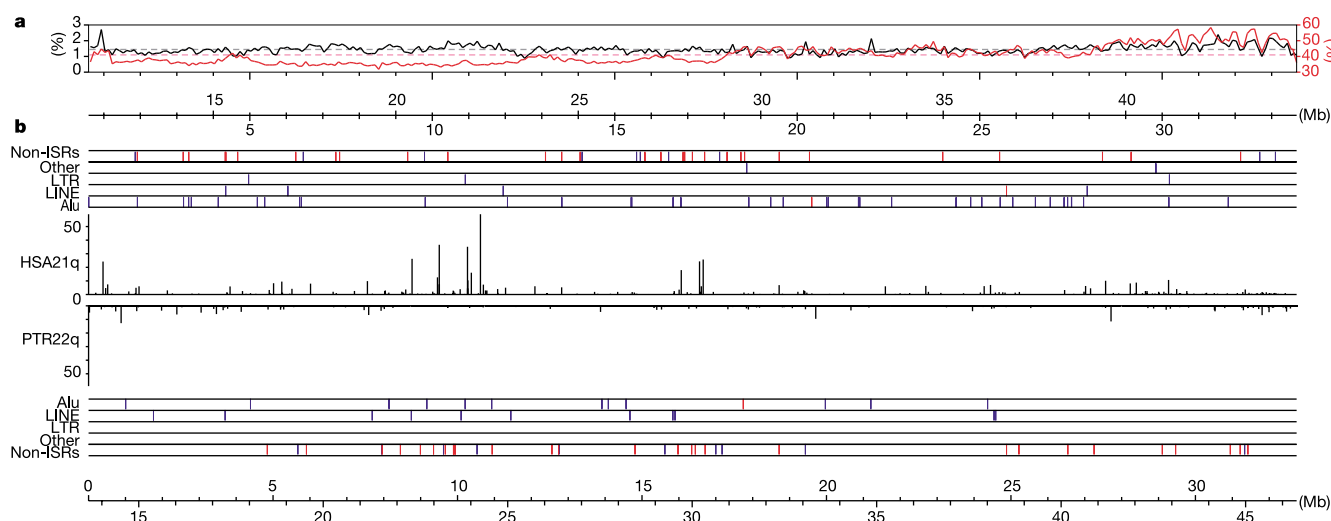


Figure 1 Overview of the differences between HSA21q and PTR22q. **a**, Nucleotide divergence level between HSA21q and PTR22q (bin size = 100 kb) is shown in black and G+C content (%) is in red. Dashed black and red lines indicate average values for divergence level and G+C content, respectively. Abscissa origin is the centromere–euchromatin boundary position in HSA21q (11,000,001 bp) and the scale is in megabases. **b**, *In silico* D-loop (a computer simulation of a hypothetical hybridization experiment using two chromosomal DNAs). The top half represents HSA21q and the

bottom half represents PTR22q. Each hypothetical loop-out (regions of DNA that do not hybridize each other) is shown as a vertical line. The y-axis indicates size of loop-outs (kb). Positions of experimentally verified lineage-specific insertions (blue) and deletions (red) are shown in five lanes along the chromosomes. The lanes, from inside to outside, represent the content of the indels: Alu elements, LINEs, LTRs, other ISRs (Other) and non-ISRs. The scale bar indicates the physical position in megabases relative to the telomere of the short arm (outside track) and to the centromere (inside track).

Lineage-specific insertions and deletions

Through alignment of the high-quality chromosomal sequences of HSA21q and PTR22q we identified about 68,000 indels in total. Greater than 99% of the indels are shorter than 300 bp, but there is a clear abundance of those around 300 bp in size (Fig. 2). These sites are probably produced either through human insertions/chimpanzee deletions or vice versa. Thus the precise identification of these molecular events in the two genomes is essential to understand the processes underlying human and chimpanzee evolution. For this purpose, we tested 567 indels larger than 300 bp using DNA samples from five human, five chimpanzee, one gorilla and two orang-utan individuals by polymerase chain reaction (PCR) amplification using the same primer sets to classify in which lineage these indels arose (see Methods and Supplementary Information). We compared the size of the successfully amplified DNA fragments from 219 indels, of which 193 showed lineage-specific changes in size. Thus, we were able to distinguish insertion from deletion events independently in human and chimpanzee lineages, and to estimate the original state of these regions in the genome of the last common ancestor.

We then classified the indels based on their contents (Fig. 1b). Insertions were mostly produced by the integration of Alu and L1 elements, whereas deletions were not related to particular repetitive structures except in a few cases. We observed different distributions of newly integrated Alu elements between HSA21q and PTR22q: 56% of new Alu elements in HSA21q are inserted in the half of the chromosome with high G+C content, whereas 70% in PTR22q are in the half with low G+C content; new LINEs are more frequent in the half with low G+C content of both chromosomes.

The plots of human insertions and chimpanzee insertions show different multimodal curves (Fig. 3). On the basis of the positions of the insertion sites on HSA21q and PTR22q, we found that most (41 and 13, respectively) of the insertions (300–350 bp in length) were members of the AluY family in both chromosomes. In contrast, only a smaller number of insertions, mostly L1 and LTR elements, were found in the 370–1,000-bp size range. Notably, human and chimpanzee deletion plots form a similar linear line, suggesting a relationship between logarithmic size of deletions and the cumulative frequency in both species (Fig. 3).

We also identified integration of L1PA2 elements after human–chimpanzee speciation, indicating that L1PA2 has been active in both human and chimpanzee lineages, although the activity seems to be lower in the human lineage. Two L1PA2 elements reside in different strands but overlap in a single inserted region in PTR22q, suggesting a single L1PA2 integration event followed by an inversion event within the same region. We also found that some insertions in PTR22q lie within Alu elements (mostly AluSx) on the same strand.

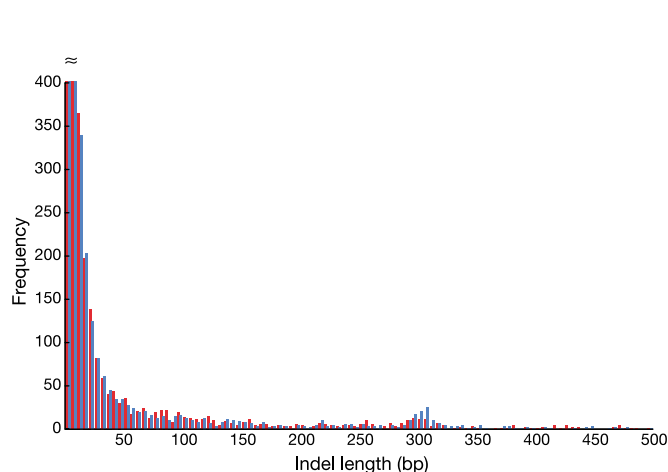


Figure 2 Size distribution of indels. All of the indels are calculated as insertions either in HSA21q (blue) or PTR22q (red). The first two bins are off the scale.

Unlike the insertions, deletions do not correspond exactly to any ISR elements, indicating that deletion events are independent of ISRs. However, one of the deleted regions in HSA21q corresponds perfectly to a single AluY element in PTR22q, and a deleted region in PTR22q corresponds almost perfectly to a single AluYb8 element in HSA21q. In the former case, there are two identical 10-mer segments around the deleted AluY element, and in the latter case, the AluYb8 element is embedded within a single AluSx element at a site in the 14-bp A-rich region in the middle of the Alu element, generating 14–15-bp poly-A/T stretches around AluYb8. Thus, the deletion of these elements may also have been generated by homologous recombination between these relatively short identical or similar flanking segments.

Calculations from the indels in the 300–5,000-bp range indicate that both chromosomes have undergone a net loss in size since speciation despite frequent insertion events: HSA21q has gained 32 kb but lost 39 kb, whereas PTR22q has gained 25 kb and lost 53 kb. This suggests that the ancestral chromosome was larger than both HSA21q and PTR22q, and that PTR22q has suffered more losses than HSA21q since speciation. The large indels (>5 kb) detected in the sequences, which were experimentally confirmed, are found in the pericentromeric, 10 Mb, 17 Mb and 29 Mb regions. HSA21q has more indels greater than 10 kb than PTR22q.

With the knowledge of which Alu family element was inserted after speciation, we carried out an evolutionary analysis of the AluY families that have been inserted into HSA21q and PTR22q. A neighbour-joining analysis revealed that such AluY elements can be largely separated into chimpanzee and human groups and suggesting contribution from a few active elements (Fig. 4). Taken as a whole, these results indicate that the expansion of particular elements was repeated several times during the course of evolution. Humans seem to have experienced such expansions more frequently and more recently than chimpanzees. If we could determine the oldest expansion event through genomic comparison, we might be able to identify whether such an Alu burst was the driving force for speciation between the two species from the common ancestor. Amplification of Alu elements during the evolution of primates and alteration of gene functions through the insertion of repetitive elements has been discussed in many previous studies^{25–34}. However, further wide-ranging analyses comparing the chimpanzee and other primate genomes is necessary to clarify these points.

Chimpanzee and human single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) provide important clues for detecting ancestral and mutant alleles within the human

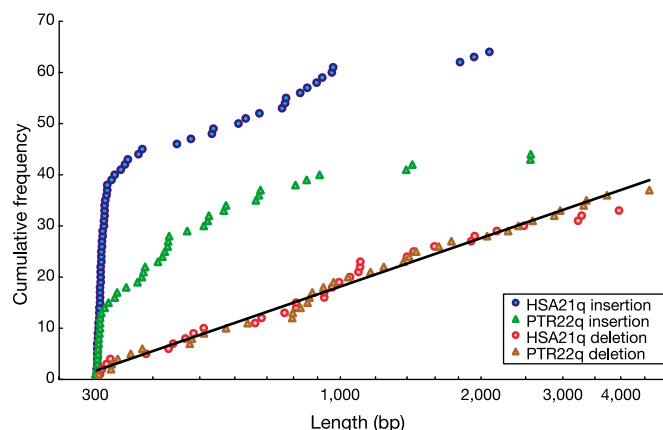


Figure 3 Size-dependency of indel frequency. Cumulative counts of experimentally determined lineage-specific insertions and deletions. The x axis represents the log-scaled size of indels (bp), whereas the y axis represents the cumulative counts of lineage-specific insertions and deletions.

population. The chimpanzee genomic sequence is the best resource for inferring the ancestral allele of any human SNP locus. We thus unambiguously reconstructed 19,985 ancestral states from 21,435 human SNP sites in HSA21q⁷.

Transitional changes are more frequent than transversions, as expected. Among transitions, G→A and C→T changes (19.6% and 20.3%, respectively)³⁵ are more frequent than A→G and T→C changes (14.0% and 15.1%, respectively). This substitution pattern is compatible with the fact that the G+C content of the human genome is lower than 50%. We conducted the same analysis for 5,781 chimpanzee SNPs obtained from overlapping BAC sequences (Supplementary Table 9A, B), and found that both A→G and T→C transitions (12.6% and 13.2%, respectively) are slightly lower than those for human.

Because SNPs have been created relatively recently during human and chimpanzee evolution (approximately 0.5–1.0 million years ago), nucleotide substitution patterns predicted from SNP data may be slightly different than those based on current G+C content (that is, 41% for both human and chimpanzee). As the estimated equilibrium G+C content for the human genome is 0.422, versus 0.405 for chimpanzee, modern humans seem to have undergone a slight increase in G+C content compared with their ancestors,

whereas chimpanzees have undergone no clear change, although the equilibrium values for humans and chimpanzees are not far from the current G+C contents (Supplementary Table 9C, D).

We then conducted an *H* test³⁶ (Supplementary Table 10), and detected 18 10-kb DNA regions as candidates of positive selection. Three known genes, *KCNE1*, *DSCR2* and *B3GALT5*, were located in these regions. *KCNE1* and *B3GALT5* were also identified as relatively rapidly evolving genes through analysis of the ratio of the number of amino acid substitutions per site to silent substitutions per site (K_A/K_S), as discussed below (see Supplementary Table 6).

Gene catalogue and characterization of coding sequences

We have annotated 284 protein-coding genes and 98 pseudogenes for HSA21q¹² (<http://chr21.molgen.mpg.de>), and 272 genes and 89 pseudogenes for PTR22q (the comparative gene catalogue of HSA21q and PTR22q, including pseudogenes, is given in Supplementary Table 3). We lacked information for six genes and 11 pseudogenes on HSA21q located partly or completely in sequencing gaps of PTR22q, and one pseudogene, *OR7E92P*, was absent in PTR22q (Supplementary Table 3). All of the conserved pseudogenes were matched in size between human and chimpanzee, except for *KRTAP21P1*, which is non-processed in HSA21q but processed in PTR22q.

Six HSA21q genes displaying hallmarks of retrogenes (see ref. 37 for review) were not found in PTR22q and were probably inserted during human evolution (or less likely, deleted during chimpanzee evolution). These are *H2BFS*, a member of the histone family S, and five members of the keratin-associated protein (KAP) gene cluster in 21q22.1. *H2BFS* was not found in the syntenic region of mouse chromosome 16, suggesting that it was inserted in HSA21q. The KAP gene cluster is too divergent in mouse to establish a conclusive comparison. Three intronless open reading frames (ORFs) have been inserted in chimpanzee PTR22q (or deleted in HSA21q). These are *HNRPA1LK1*, a heterogeneous nuclear ribonucleoprotein, *RPLPILK1*, a ribosomal protein, and *FAM28ALK1*, a gene of unknown function; all three ORFs are absent in the mouse syntenic region. All but one of the ORFs found in either human insertions/chimpanzee deletions or human deletions/chimpanzee insertions are intronless and probably represent retrogenes.

We found a ribosomal protein pseudogene (*RPL13AP*) on HSA21q with an intact ORF (*RPL13ALK1*) in PTR22q, allowing for possible functionality. Four HSA21q coding sequences (*C21orf81* and three intronless genes: *C21orf115*, *C21orf104* and *C21orf19*) have interrupted ORFs in PTR22q, precluding their functionality, and these were annotated as pseudogenes in chimpanzee (Supplementary Table 3).

All other HSA21q genes are potentially active in PTR22q, even though several genes show human-specific transcriptional isoforms due to alteration or inactivation of some of the transcript isoforms in chimpanzee (see below). The minimum nucleotide sequence identity is 83% (*KRTAP6-3*). Of the 272 annotated chimpanzee genes, we compared the human and chimpanzee coding sequences in 231 genes for which we could define a non-ambiguous ORF in both species. We omitted for that comparison the 41 entries of the chimpanzee catalogue for which we had no ORF in one of the two species (for example, ambiguous ORFs associated to short ESTs containing mostly untranslated regions (UTRs)) and genes corresponding to pseudogenes in the other species. Among the 231 genes associated to a canonical ORF, 179 show a coding sequence of identical length in human and chimpanzee and exhibit similar intron–exon boundaries. For those 179 genes, the average nucleotide and amino acid identity in the coding region is 99.29% and 99.18%, respectively. Of these, 39 genes show an identical amino acid sequence between human and chimpanzee, including seven in which the nucleotide sequence of the coding region is also identical (Supplementary Table 3). Examples of biological processes that are perfectly conserved involve transcriptional regulators (*RUNX1*,

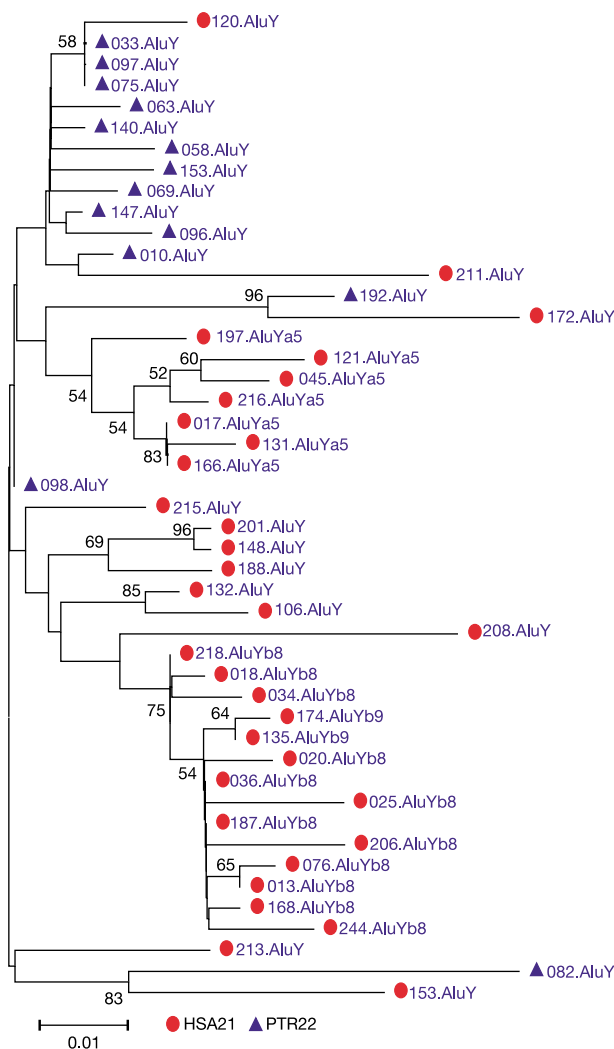


Figure 4 Evolutionary relationships among young AluY elements. Neighbour-joining tree of all AluY families that have been inserted into HSA21q (red circles) or PTR22q (blue triangles) after speciation. Bootstrap proportions greater than 50% are indicated at the corresponding branches (1,000 replicates). The scale indicates the evolutionary distance of 0.01.

PKNOX1, *ERG*, *GABPA*, *U2AF1*), metabolic enzymes (*ATP5O*, *CBRI*, *PFKL*, *CRYZL1*, *SOD1*, *ABCG1*), gene products associated with signal transduction generally showing patterned expression during development (*GRIK1*, *CXADR*, *DSCR5*, *PCP4*, *DSCAM*, *DYRK1A*, *S100B*, *C21orf4*) and gene products involved in protein folding and degradation (*SMT3H1*, *UBE2G2*)³⁸. A total of 140 of these 179 genes show amino acid replacements, but no gross structural changes are expected.

In contrast, 47 PTR22q genes show significant structural changes affecting at least one of their transcript isoforms. Fifteen genes have indels within their coding region yet retain frame consistency in all but one case (*TCP10L*) (Supplementary Table 4). Marked changes are observed in *PCNT2*, a component of the filamentous matrix of the centrosome initiating the nucleation of spindle microtubules, and in *TCP10L*, a t-complex protein. The third exon of *PCNT2* is shorter in chimpanzee owing to a deletion of 195 bp corresponding to five highly similar copies of a 13-mer repeating unit. Human *PCNT2* has seven repeats, the orthologous chimpanzee gene has only two whereas the mouse has none, suggesting that the repeats were inserted during primate evolution. *TCP10L* has a deletion of 17 nucleotides within the fourth exon in chimpanzee. This generates a transcript that uses the last 16 nucleotides of exon 4 and the adjacent 23 nucleotides that are intronic in human, the gene product of which is predicted to have a short frame shift in the middle of the protein (Supplementary Table 4). Thirty-two genes show changes modifying either the first ATG or the stop codon in at least one of their associated transcripts (Supplementary Table 5).

Five chimpanzee genes could not be classified because they displayed structural changes caused by indels (*SH3BGR*, *SYNJ1*, *C21orf96* and *TMPRSS3*) or a substitution in the ATG codon (*C21orf18*). These changes correspond to polymorphisms in human and may not be specific to chimpanzee.

Our data suggest that indels within coding regions represent one of the major mechanisms generating protein diversity and shaping higher primate species. We observed an additional level of functional diversity generated through the occurrence of multiple alternative transcripts for a single gene, some of the isoforms being structurally modified or not functional in chimpanzee. We are currently lacking cDNA information to define precisely the structure of these genes but we provide here a framework allowing for experimental verification of the transcript isoforms.

Taken together, gross structural changes affecting gene products are far more common than previously estimated (20.3% of the PTR22 proteins, as listed in Supplementary Tables 4 and 5). In addition, 87 genes in the catalogue show mutations in at least one of the splice sites. However, in all cases the modified sequence still fits the consensus sequence and is therefore expected to be functional. More subtle effects, such as changes in transcript stability or processing, cannot be ruled out.

K_A/K_S analysis

The neutrality of sequence differences found between orthologous pairs of human and chimpanzee genes can be assessed using K_A and K_S values ($K_A/K_S \cong 1$ indicates neutral evolution; $K_A/K_S > 1$ indicates positive selection, and $K_A/K_S < 1$ indicates negative selection; that is, purifying selection). K_A , K_S and other related values were calculated for 231 genes on PTR22q, with 10% of the genes having a K_A/K_S ratio > 1 , the highest value being 3.37 for the human hair keratin-associated protein 23-1 gene (*KRTAP23-1*; data are summarized in Supplementary Table 6), although these values were not statistically confirmed (Supplementary Fig. 1).

Relatively rapidly evolving genes may be estimated from K_A , $K_A + K_S$, or just nucleotide divergence values. Three KAP genes, *KCNE1* (human cardiac delayed rectifier potassium channel protein), *TCP10L* (t-complex protein 10A-2), *B3GALT5* (UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5), *IGSF5* (immunoglobulin superfamily-5 like) and several other

genes are found in this category (Supplementary Table 6). Genes showing statistically significant low K_A/K_S values are thought to be evolving under purifying selection; for example, *C21orf113*, *PFKL*, *AIRE*, *ITGB2*, *TMPRSS3* and *AGPAT3*. *PFKL* has a K_A/K_S ratio of 0, and is likely to be under strong purifying selection over the entire gene region.

Recently, Clarke *et al.*³⁹ reported a similar analysis against 7,645 chimpanzee gene sequences including 70 PTR22q gene sequences, in which *KRTAP23-1* was not included. Furthermore, we found discrepancies with a previous analysis reporting K_A/K_S values for some of the HSA21q genes⁴⁰. One explanation is that the previous work used only PCR-amplifiable exons for the analysis.

Comparative gene expression analysis

Using Affymetrix HG U95 arrays, we compared the gene expression profiles of HSA21q genes between humans and chimpanzees in two tissues: 202 genes in brain (cingulate cortex, HG U95Av2, B, C, D and E) and 96 genes in liver (HG U95Av2). We detected 60 genes expressed in brain (this study) and 40 in liver⁴¹ in at least one species. Of these, 9 in the brain and 12 in the liver showed a significant change in expression level between humans and chimpanzees in the range of a 1.5–10-fold difference (Supplementary Table 7). Overall, the proportion of genes showing changes in expression level on HSA21 was not significantly different from the rest of the genome (liver: $\chi^2 = 1.87$, $P = 0.172$; brain: $\chi^2 = 1.12$, $P = 0.290$)⁴¹.

We explored whether the sequence divergence in the different gene regions can predict how a gene is differently expressed between the two species using a Mann–Whitney *U*-test. The data show that there is a trend for regions with a high divergence in the 5' UTR to differ in messenger RNA levels between humans and chimpanzees ($N = 18$, 47; $P = 0.043$). Notably, genes in which the sequence divergence of an associated CpG islands is high are also more likely to have changed their expression ($N = 13$, 39; $P = 0.046$). However, as CpG islands often stretch into the first exon, the correlations of CpG island and 5' UTR divergence with expression changes are not independent. In contrast, we found no significant association between the divergence of non-degenerate sites, 3' UTRs, intergenic and intronic regions and variation in expression levels (Supplementary Table 8). It has been proposed that 5' UTRs in humans might have been under positive selection⁴², possibly due to their involvement in the regulation of gene expression levels.

Some of the genes displaying significant changes in protein sequence or differences in expression between human and chimpanzee might be correlated with physiological or disease susceptibility differences exhibited between the two species⁴³. For instance, *IFNAR2*, *IFNGR2*, *CXADR*, *ITSN1* and *CRYZL1* are directly or indirectly involved in the immune response against various pathogens. *SH3BGR* is strongly expressed in the developing heart, *C21orf2* is expressed in the peripheral nervous system, *SYNJ1* and *ANKRD3* are signalling molecules acting in early brain development, *MCM3AP* is associated with cell cycle progression, *ETS2* is a transcription factor essential for embryonic development and *COL18A1* is a collagen gene that is mutated in human Knobloch syndrome associated with encephalocele⁴⁴.

Promoter analysis

We analysed the upstream region of genes that showed significant expression changes in liver and brain between human and chimpanzee, as well as the upstream region of each corresponding mouse gene. Computational analysis of the transcription-factor-binding sites within the 1-kb upstream region of each gene is summarized in Supplementary Table 11. Transcription-factor-binding sites were compared in the three species, and those specific to either human or chimpanzee were found in most genes. All of the specific transcription-factor-binding sites were caused by base substitutions in either human or chimpanzee, but these may not clearly account for the expression changes observed in this study. To assess precisely the

expression changes, further analysis of promoter regions as well as other factors, such as enhancers and suppressors located outside the promoter region and the effects of 5' and 3' UTRs on mRNA stability, are needed.

Conclusion

This study shows a chromosome-wide comparison between human and chimpanzee based on high-quality sequences, and provides the first integrated picture of genetic changes during human evolution. The data presented here suggest that the biological consequences due to the genetic differences are much more complicated than previously speculated. We hope that our work offers a framework for the design of future studies to examine differences between the two species. □

Methods

Mapping, sequencing and data availability

The details concerning mapping and sequencing are summarized in Supplementary Tables 1 and 2 (see also <http://chimp22pub.gsc.riken.jp>). Briefly, three BAC libraries—PTB1 (ref. 3), RPCI-43 and CHORI-251 (<http://bacpac.chori.org>), constructed from three male individuals—and a chimpanzee chromosome 22 fosmid library—PTF22 (12-fold coverage)—were used to isolate clones for this analysis. The first set of the seed clones was selected from BAC end data⁷ and PCR screening of expanded BAC libraries using high-density human STS primers placed at roughly 20-kb intervals. Only the clones having multiple STS markers and sharing common STSs with neighbouring clones were considered as a member of a clone contig. These clones were then subjected to partial sequencing, from which new chromosome-walking primers were designed for further screening. Clone overlaps were then examined before choosing a set of minimum-tiling-path clones for full-scale high-quality sequencing. These steps were repeated until all gaps were closed or no additional clones could be identified. Some representative clones in contigs were also examined cytogenetically to confirm their localization on the particular chromosome; however, not all such clones were included in the minimum tiling path. Most of the pericentromeric and subtelomeric regions are not included in this study because of the difficulty in identifying clones with good sequence matches. Two PTR22 clones, PTB-242K04 and CH251-010A09, have a terminal region that extends into one of the three clone gaps in HSA21q by 11,546 bp in total. Another clone, PTB-190I13, spans another HSA21q gap with a 9,284-bp region for the gap where the G+C content is high (54.2%), and a sequence gap of 1,165 bp in length still remains.

All of the clone data have been released from DDBJ/EMBL/GenBank (accession numbers are listed in Supplementary Table 1).

Alignment between HSA21q and PTR22q and contig construction

Each PTR22 clone sequence was aligned to the HSA21q data using NCBI BLAST2. From the BLAST hits, we chose the best match for each site of the corresponding region in HSA21q. Details of the alignment procedure can be seen in Supplementary Information. The clone contig was generated by aligning the overlapping regions of each neighbouring clone. We chose to include the sequence of the clone for the overlap with which the alignment with the counterpart region in HSA21q showed a lower divergence level. On the basis of this clone contig map, the region represented in the contig is extracted from each clone alignment to produce the whole chromosome alignment. To fix any small inconsistencies around the clone boundaries, the final alignment was checked manually. The nucleotide divergence level was calculated from the regions that aligned best with HSA21q.

Detection of lineage-specific insertions and deletions

We selected 567 apparent insertions ≥ 300 bp from both the human and chimpanzee sequences and designed PCR primers from their flanking sequences. Using genomic DNA samples from five chimpanzees, five humans, one gorilla and two orang-utans as a template, the size of the PCR products was examined. After amplification, reaction products were separated through 1% agarose gel electrophoresis for comparison of the product sizes. We used only the indels that showed a significant size difference between chimpanzee and human, and one of them was of equivalent size to gorilla and orang-utan, for the analysis (*t*-test).

SNP analysis

The HSA21q SNPs⁷ in dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) were compared to the corresponding PTR22q sites to infer the ancestral states of the two human alleles. We developed a modified version of the *H* test³⁶ based on the computer program at <http://crimp.lbl.gov/Htest.html> so as to detect positive selection disregarding singletons (details in Supplementary Information). The pattern of nucleotide substitutions was also estimated using the human SNP data and chimpanzee BAC sequence overlap data generated in the present study. We estimated the substitution pattern³⁵ and equilibrium frequencies from the nucleotide substitution matrix^{45,46}.

Genomic annotation

Protein-coding genes in the PTR22q sequence were annotated by extracting the orthologous regions on HSA21q (in the gene catalogue) from the genomic sequence. In addition, PTR22 sequences with no match to the HSA21 gene catalogue were blasted

against the complete nr-db, and cDNA matches were retained as potential PTR22-specific genes. Gene orthology with mouse, rat, zebrafish, pufferfish and *Ciona* was calculated from the HSA21 genes, and conceptual translations and multiple alignments were constructed using CLUSTALW. Detailed information is provided in the Supplementary Information.

Comparative gene expression analysis

For all arrays only the oligonucleotides that matched perfectly between the human and chimpanzee sequences were used for analysis. Gene expression levels were compared separately for brain and liver in all nine possible pairwise comparisons between the three individuals of each species.

Received 16 February; accepted 14 April 2004; doi:10.1038/nature02564.

- King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
- Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
- Fujiyama, A. *et al.* Construction and analysis of a human–chimpanzee comparative clone map. *Science* **295**, 131–134 (2002).
- Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4**, 20–28 (2003).
- Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
- Hacia, J. G. *et al.* Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nature Genet.* **18**, 155–158 (1998).
- Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
- Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99**, 13633–13635 (2002).
- Nickerson, E. & Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50**, 368–372 (1998).
- Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
- Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Frazer, K. A. *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11**, 1651–1659 (2001).
- Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13**, 341–346 (2003).
- McClure, H. M., Belden, K. H., Pieper, W. A. & Jacobson, C. B. Autosomal trisomy in a chimpanzee: resemblance to Down's syndrome. *Science* **165**, 1010–1012 (1969).
- Ewing, B., Hillier, L., Wendt, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
- Orti, R. *et al.* Conservation of pericentromeric duplications of a 200-kb part 25 of the human 21q22.1 region in primates. *Cytogenet. Cell Genet.* **83**, 262–265 (1998).
- Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490–1497 (2002).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Sawada, I. *et al.* Evolution of Alu family repeats since the divergence of human and chimpanzee. *J. Mol. Evol.* **22**, 316–322 (1985).
- Myers, J. S. *et al.* A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71**, 312–326 (2002).
- Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**, 358–368 (2003).
- Kim, M., Carman, C. V. & Springer, T. A. Bidirectional transmembrane signaling by cytoplasmic domain separation in integrins. *Science* **301**, 1720–1725 (2003).
- Roy, A. M. *et al.* Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**, 1485–1495 (2000).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. Sources and evolution of human Alu repeated sequences. *Proc. Natl Acad. Sci. USA* **85**, 4770–4774 (1988).
- Batzer, M. A. & Deininger, P. L. A human-specific subfamily of Alu sequences. *Genomics* **9**, 481–487 (1991).
- Leefflang, E. P., Chesnokov, I. N. & Schmid, C. W. Mobility of short interspersed repeats within the chimpanzee lineage. *J. Mol. Evol.* **37**, 566–572 (1993).
- Zietkiewicz, E., Richer, C., Makalowski, W., Jurka, J. & Labuda, D. A young Alu subfamily amplified independently in human and African great apes lineages. *Nucleic Acids Res.* **22**, 5608–5612 (1994).
- Batzer, M. A. *et al.* Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247**, 418–427 (1995).
- Roy, A. M. *et al.* Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**, 149–161 (1999).
- Chou, H. H. *et al.* Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc. Natl Acad. Sci. USA* **99**, 11736–11741 (2002).
- Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316**, 1033–1040 (2002).
- Deininger, P. L. & Batzer, M. A. Mammalian retroelements. *Genome Res.* **12**, 1455–1465 (2002).
- Gojobori, T., Li, W. H. & Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360–369 (1982).
- Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
- Brosius, J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**, 115–134 (1999).

38. The HSA21 expression map initiative. A gene expression map of human chromosome 21 orthologues in the mouse. *Nature* **420**, 586–590 (2002).
39. Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
40. Shi, J. *et al.* Divergence of the genes on human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. *Proc. Natl Acad. Sci. USA* **100**, 8331–8336 (2003).
41. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
42. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
43. Varki, A. A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10**, 1065–1070 (2000).
44. Suzuki, O. T. *et al.* Molecular analysis of collagen XVIII reveals novel mutations, presence of a third isoform, and possible genetic heterogeneity in Knobloch syndrome. *Am. J. Hum. Genet.* **71**, 1320–1329 (2002).
45. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *A Model of Evolutionary Change in Proteins* (ed. Dayhoff, M. O.) (Natl Biomed. Res. Found., Silver Springs, Maryland, 1978).
46. Tajima, F. & Nei, M. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J. Mol. Evol.* **18**, 115–120 (1982).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to T. Ito, C. Kawagoe, T. Kojima, X. Son, A. Beck, K. Borzym, S. Gelling, V. Gimmel, K. Heitmann, A. Kel, S. Klages, N. Lang, I. Mueller, M. Sontag, R. Yildirimman, J. Wickings, C. Baumgart, O. Mueller, T. T. Liao, H. Tsai, Y. Huang, Y. Liu and all the technical staff of the contributing genome centres. This work was supported in part by a Special Fund for RIKEN Genomic Sciences Center and Grant-in-Aid for Scientific Research on Priority Areas ‘Genome Science’ from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Ministry of Education and Research, Germany; the Chinese International Science and Technology Cooperation Project, Ministry of Science and Technology, China; the Chinese High-Tech Research and Development Program, Shanghai Commission for Science and Technology; the National Research Program for Genomic Medicine of National Science Council, Taiwan; and the Ministry of Science and Technology, Korea.

Authors' contributions H. Watanabe, A. Fujiyama, M. Hattori, N. Saitou, H.-S. Park, S.-Y. Wang, M.-L. Yaspo and Y. Sakaki contributed to the consortium leadership.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.S. (sakaki@gsc.riken.jp). The accession numbers of the BAC clone sequences have been deposited in DDBJ/EMBL/GenBank and can be found in Supplementary Table 1.

H. Watanabe^{1,2*}, A. Fujiyama^{1,3}, M. Hattori^{1,4}, T. D. Taylor¹, A. Toyoda¹, Y. Kuroki¹, H. Noguchi¹, A. BenKahla⁵, H. Lehrach⁵, R. Sudbrak⁵, M. Kube⁵, S. Taenzer⁶, P. Galgoczy⁶, M. Platzer⁶, M. Scharfe⁷, G. Nordtsiek⁷, H. Blöcker⁷, I. Hellmann⁸, P. Khaitovich⁸, S. Pääbo⁸, R. Reinhardt⁵, H.-J. Zheng⁹, X.-L. Zhang⁹, G.-F. Zhu⁹, B.-F. Wang⁹, G. Fu⁹, S.-X. Ren⁹, G.-P. Zhao⁹, Z. Chen^{9,10}, Y.-S. Lee¹¹, J.-E. Cheong¹¹, S.-H. Choi¹¹, K.-M. Wu¹², T.-T. Liu¹³, K.-J. Hsiao^{13,14}, S.-F. Tsai^{12,13}, C.-G. Kim¹⁵, S. Oota¹⁵, T. Kitano¹⁵, Y. Kohara¹⁵, N. Saitou¹⁵, H.-S. Park¹¹, S.-Y. Wang⁹, M.-L. Yaspo⁵ & Y. Sakaki¹

Affiliations for authors: 1, RIKEN, Genomic Sciences Center, Yokohama 230-0045, Japan; 2, Nara Institute of Science and Technology, Ikoma 630-0101, Japan; 3, National Institute of Informatics, Tokyo 101-8430, Japan; 4, Kitasato University, Sagami-hara 228-8555, Japan; 5, Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin-Dahlem, Germany; 6, Institut für Molekulare Biotechnologie, Genomanalyse, D-07745 Jena, Germany; 7, German Research Centre for Biotechnology (GBF), D-38124 Braunschweig, Germany; 8, Max-Planck-Institut für Evolutionäre Anthropologie, D-04103 Leipzig, Germany; 9, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China; 10, State Key Laboratory of Medical Genomics, Rui Jin Hospital, Shanghai 200025, China; 11, Genome Research Center, Korea Research Institute of Basic and Biotechnology, Daejeon 305-333, Korea; 12, National Health Research Institutes, Taipei 115, Taiwan; 13, National Yang-Ming University, Taipei 112, Taiwan; 14, Taipei Veterans' General Hospital, Taipei 112, Taiwan; 15, National Institute of Genetics, Mishima 411-8540, Japan

*Present address: Hokkaido University, Sapporo 060-0814, Japan

Electron-hole symmetry in a semiconducting carbon nanotube quantum dot

Pablo Jarillo-Herrero^{1,2}, Sami Sapmaz¹, Cees Dekker¹, Leo P. Kouwenhoven^{1,2} & Herre S. J. van der Zant¹

¹Kavli Institute of Nanoscience, and ²ERATO project on Mesoscopic Correlations, Delft University of Technology, PO Box 5046, 2600 GA, Delft, The Netherlands

Optical and electronic phenomena in solids arise from the behaviour of electrons and holes (unoccupied states in a filled electron sea). Electron-hole symmetry can often be invoked as a simplifying description, which states that electrons with energy above the Fermi sea behave the same as holes below the Fermi energy. In semiconductors, however, electron-hole symmetry is generally absent, because the energy-band structure of the conduction band differs from the valence band¹. Here we report on measurements of the discrete, quantized-energy spectrum of electrons and holes in a semiconducting carbon nanotube². By applying a voltage to a gate electrode, an individual nanotube is filled controllably with a precise number of either electrons or holes, starting from one. The discrete excitation spectrum for a nanotube with N holes is strikingly similar to the corresponding spectrum for N electrons. This observation of near-perfect electron-hole symmetry³ demonstrates that a semiconducting nanotube can be free of charged impurities, even in the limit of few electrons or holes. We furthermore find an anomalously small Zeeman spin splitting and an excitation spectrum indicating strong electron-electron interactions.

Carbon nanotubes can be metallic or semiconducting depending on their chirality. Electron transport through individual nanotubes has been studied for both classes². Nanotubes of finite length have a discrete energy spectrum. Analogous to studies on semiconducting quantum dots, these discrete states can be filled with electrons, one by one, by means of a voltage applied to a nearby gate electrode⁴. Whereas metallic nanotubes have shown clean quantum dot (QD) behaviour^{5–7}, this has not been achieved in semiconducting single-wall nanotubes (SWNTs). Theory indicates that semiconducting tubes are more susceptible to disorder than metallic ones^{8,9}. Disorder typically divides a semiconducting nanotube into multiple islands, preventing the formation of a single, well-defined QD. Consequently, the electronic spectrum of semiconducting SWNTs has not been resolved before.

We report here on clean semiconducting tubes, and focus on the regime of a few charge carriers (electrons or holes). We use high-purity carbon nanotubes (HiPco, ref. 10), which are deposited with low density on a doped Si substrate (serving as a backgate) that has an insulating SiO₂ top layer^{11,12}. Individual nanotubes are electrically contacted with source and drain electrodes (50 nm Au on 5 nm Cr). We then suspend the nanotubes by etching away part of the SiO₂ surface¹². We generally find that removing the nearby oxide reduces the amount of potential fluctuations (that is, disorder) in the nanotubes, as deduced from transport characteristics.

In this paper we focus on one particular semiconducting device that shows regular single QD behaviour for both few-hole and few-electron doping. The distance between the electrodes in this device is 270 nm (Fig. 1a, b). The dependence of the linear conductance on gate voltage shown in Fig. 1c is typical for semiconducting p- and n-type behaviour^{13,14}. A low-temperature measurement around zero gate voltage (Fig. 1d) shows a large zero-current gap of about 300 meV in bias voltage, reflecting the semiconducting character of this nanotube. The zigzag pattern outside the semiconducting gap is due to Coulomb blockade⁴. These Coulomb blockade features are

more evident in Fig. 1e, where a high-resolution measurement of the differential conductance shows the semiconducting gap with the first two adjacent Coulomb blockade diamonds.

The identification of the Coulomb diamonds for the first electron and first hole allows for an unambiguous determination of the particle number as we continue to fill the QD by further changing the gate voltage. Figure 2a shows the filling of holes, one by one, up to 20 holes. The region for the first two holes is enlarged in Fig. 2b. The regularity in the Coulomb diamonds indicates a nanotube that is free of disorder. A closer inspection shows that the size of the Coulomb diamonds varies periodically on a smooth background as the hole number increases (Fig. 2c). The alternating, even-odd pattern in this addition energy, E_{add} , reflects the subsequent filling of discrete orbital states with two holes of opposite spin⁴.

We first focus on the additional discrete lines outside the Coulomb diamonds running parallel to its edges, as for instance indicated by arrows in Fig. 2b. Whereas the upper-left edge of the N -hole diamond reflects the ground-state energy of the $(N+1)$ hole, the extra lines located at higher voltages, V , represent the discrete excitation spectrum for $(N+1)$ -holes⁴. The spacing in V directly measures the energy separation between the excitations. Such discrete spectra have not been obtained before for semiconducting nanotubes.

We now compare the excitation spectra for a particular hole (h) number with the same electron (e) number. The left and right columns in Fig. 3 show the spectra for, respectively, holes and electrons. The upper row compares the spectra for 1 h and 1 e. The

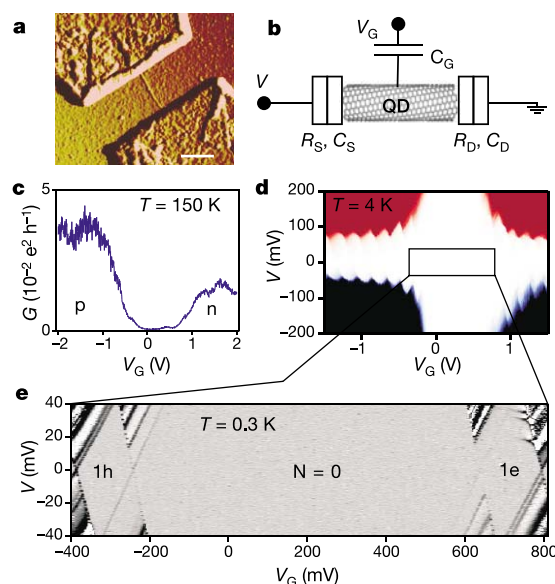


Figure 1 Sample and characterization. **a**, Atomic force microscope image of the device before suspension (scale bar, 200 nm). **b**, Device scheme. The nanotube QD is connected to source and drain electrodes by tunnel barriers characterized by resistances R_S , R_D and capacitances C_S , C_D . The backgate is represented by a capacitor C_G . The direct-current source-drain current, I , is recorded in the measurements as a function of source-drain voltage V and gate voltage V_G . Current-voltage (I - V) characteristics are numerically differentiated to obtain the differential conductance, dI/dV . **c**, Linear conductance, G , as a function of V_G at a temperature, $T \approx 150$ K showing the p- and n-type conducting regions separated by the semiconducting gap. **d**, Large-scale plot of the current (blue, negative; red, positive; white, zero) versus both V and V_G at $T = 4$ K. **e**, High-resolution measurement of the differential conductance as a function of V and V_G in the central region of **d** at 0.3 K. Between $V_G \approx -250$ and 650 mV, the nanotube QD is depleted entirely from mobile charge carriers. As V_G increases (decreases), one electron (hole) enters the dot as indicated in the right (left) Coulomb diamond.

yellow arrows in Fig. 3a point at the first three excited states for a single hole. (Note that only lines with positive slopes are observed, because of asymmetric tunnel barriers⁴.) Yellow arrows in Fig. 3b indicate the corresponding first three excitations for a single electron. (Figure 4 explains this correspondence.) Remarkably, we have simply mirror-imaged the arrows from the hole to the electron side without any adjustment of their spacing. We thus find that the 1h and 1e excitations occur at the same energy. Because one-particle systems are free from particle–particle interactions, this symmetry implies that the confinement potential for electrons is the same as for holes.

Electron–hole symmetry also survives interactions, as demonstrated in the lower rows in Fig. 3. Again, the arrows pointing at the hole excitations have simply been mirror-imaged to the electron side. Thus, we indeed find that the spectra for 2h and 2e and for 3h and 3e show virtually perfect electron–hole symmetry in the excitation spectra. Taking a closer look, we can see that the relative intensities of the excitation lines also display electron–hole symmetry.

The quality of our data allows for a quantitative analysis. The addition energy is defined as the change in electrochemical potential when adding the $(N + 1)$ charge to a QD already containing N charges. The constant-interaction model⁴ gives $E_{\text{add}} = U + \Delta E$, where $U = e^2/C$ is the charging energy ($C = C_S + C_D + C_G$; where C is the total capacitance and S , D and G refer to source, drain and gate, respectively) and ΔE is the orbital energy difference between $N + 1$ and N particles on the QD. In the case of a semiconductor QD, the addition energy for adding the first electron to the conduction band equals $U + E_{\text{gap}}$. From the observed gap size of 300 meV and $U \approx 50$ meV, we determine the semiconducting gap $E_{\text{gap}} \approx 250$ meV, which corresponds to a nanotube diameter of 2.7 nm (ref. 3). Atomic force microscopy measurements, which

usually underestimate the real height¹⁵, indicate an apparent tube height of 1.7 ± 0.5 nm.

Because two electrons with opposite spin can occupy a single orbital state, the constant-interaction model predicts an alternating value for E_{add} , where $E_{\text{add}} = U$ for $N = \text{odd}$, and $E_{\text{add}} = U + \Delta E$ for $N = \text{even}$. We do indeed observe such an even–odd alternation in Fig. 2c with average $\Delta E \approx 4.3$ meV throughout the entire range of $N = 1$ to 30. Measurements of the Zeeman spin-splitting in a magnetic field (see Supplementary Information) confirm our assignment of even–odd particle number: lines corresponding to ground states for odd N split (that is, total spin = $1/2$), whereas even- N lines do not split (that is, total spin = 0). Figure 2e shows the value of the Zeeman energy for the one-hole orbital states as a function of magnetic field. The data yield a reduced g -factor, $g \approx 1.1$, which is significantly lower than the value $g = 2$ reported on metallic nanotubes^{5,7}. (Some experiments on metallic nanotubes report deviations¹⁶.) The reduction in g -factor disappears when adding holes. The inset to Fig. 2e shows that already for nine holes the normal value is almost recovered. Lower g -factors are generally due to spin–orbit coupling, but this effect is small for carbon. It may hint at strong electron–electron interactions in the one-dimensional QD (see discussion below).

The addition energy spectrum indicates $\Delta E \approx 4.3$ meV for consecutive states as we fill the QD with holes. Previous spectra from metallic nanotubes have been analysed by considering a hard-wall potential in the nanotube, with an effective mass determined by the band structure. Our data show that this approach is not justified for semiconducting nanotubes. Lack of effective screening in one dimension and the low number of mobile charges yield a gradual potential decay from the contacts¹⁷. We have computed the addition energy spectrum for a semiconducting nanotube with a gap of ~ 250 meV for two situations (Fig. 2d): hard-wall and harmonic

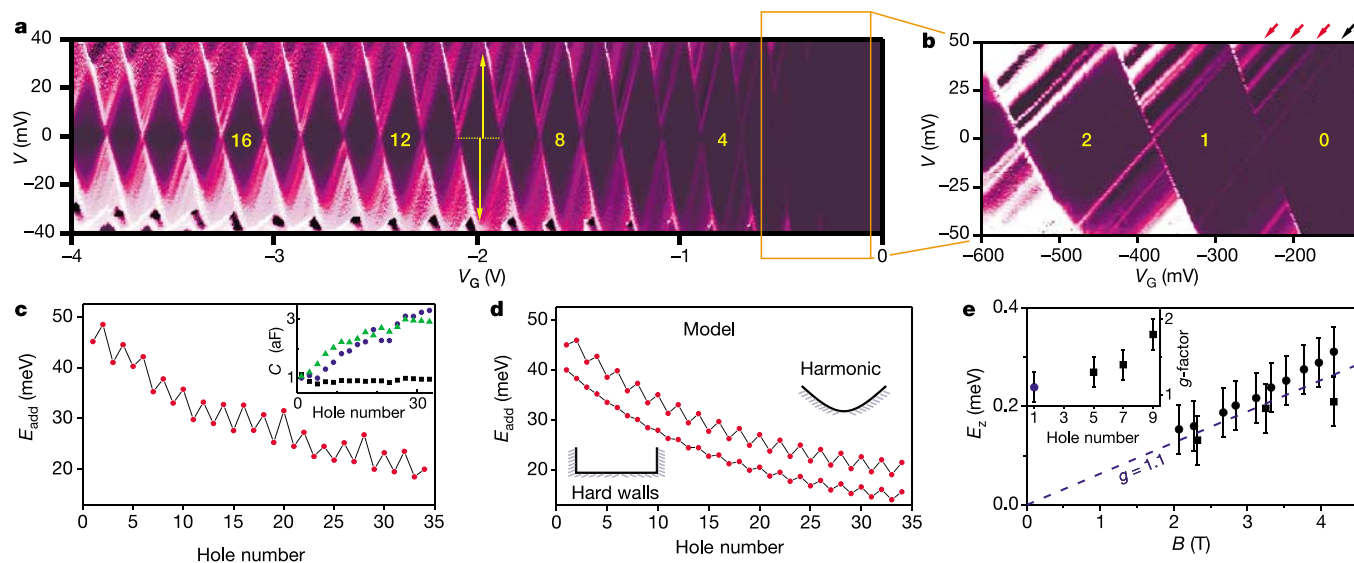


Figure 2 Few-hole semiconducting nanotube. **a**, Two-dimensional colour plot of the differential conductance, dI/dV , versus V and negative V_G at $T = 4$ K (black is zero, white is $3 \mu\text{S}$). In the black diamond-shaped regions, the number of holes (indicated) is fixed by Coulomb blockade. **b**, Zoom-in, taken at 0.3 K of the region with 0, 1 and 2 holes (white represents $dI/dV > 10$ nS). Lines outside the diamonds running parallel to the edges correspond to discrete energy excitations (the black arrow points at the one-electron ground state; the red arrows at the one-electron excited states). **c**, Addition energy, E_{add} , as a function of hole number. E_{add} is deduced from the diamond size for positive and negative V (that is, half the sum of the yellow arrows in **a**). Inset, the capacitances C_S (green), C_D (blue) and C_G (black) versus hole number. **d**, Calculation of the addition energy spectrum for a semiconducting nanotube (as an example we have taken a

zigzag (35,0), with $E_{\text{gap}} \approx 259$ meV, $m_{\text{eff}} = 0.037 m_e$; ref. 3) for a harmonic potential (top) and a hard-wall potential (bottom). The parameters for the harmonic potential are: $V(x = \pm 135 \text{ nm}) = E_{\text{gap}}/2$, where x is the distance from the centre of the nanotube (see Supplementary Information). **e**, Zeeman splitting energy, E_Z , versus magnetic field, B , for the one-hole orbital states. The data result from two different types of measurements: (1) individual gate voltage traces at fixed bias (circles) and (2) stability diagrams (squares, see also Supplementary Information). Inset, g -factor as a function of hole number. The point for $N = 1$ is the average of the data in Fig. 2e. The points for $N = 5, 7$ and 9 are obtained from co-tunnelling (see Supplementary Information).

potential of height $E_{\text{gap}}/2$ at the contacts¹⁷. For hard walls, the level spacing increases slowly up to ~ 1.9 meV for $N = 34$. In the case of a harmonic potential, the level spacing is constant, as in the experiment, and equals 2.7 meV, in reasonable agreement with the experimental value ~ 4.3 meV (see Supplementary Information).

On top of the predicted even–odd pattern, there is a monotonic decrease of the average charging energy with N , implying that C is changing. We have performed a detailed analysis of the QD electrostatics following ref. 18. The result is given in the inset to Fig. 2c. It shows that the change in C is due mainly to an increase in C_S and C_D . This increase can be assigned to a decrease of the tunnel barrier widths as the gate voltage, $|V_G|$, increases, consistent with the simultaneous decrease of dI/dV in Fig. 2b. Indeed, dI/dV

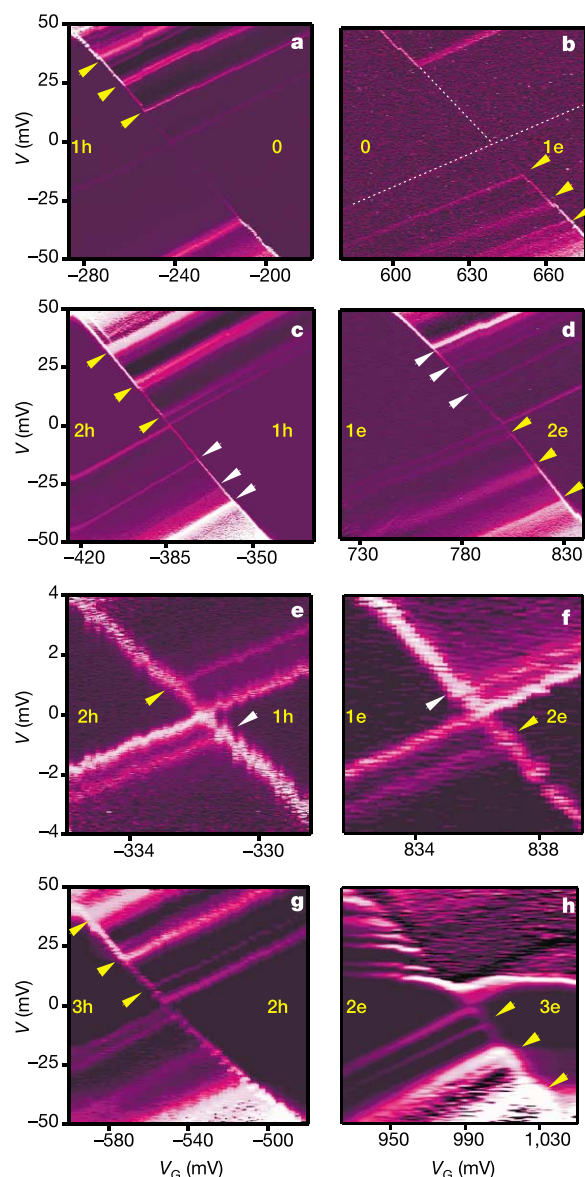


Figure 3 Excitation spectra for different electron and hole numbers demonstrating electron–hole symmetry. dI/dV is plotted versus (V, V_G) at $T = 0.3$ K. **a**, The transition from the 0 to 1 h Coulomb diamonds. **b**, Corresponding transition from 0 to 1 e. The white dotted lines in **b** are guides to the eye to indicate the diamond edge (not visible for this choice of contrast). **c** and **d**, Same for 1–2 h and 1–2 e. **e** and **f**, Low-bias zoom-in of the 1–2 h and 1–2 e crossings. **g** and **h**, Crossings corresponding to the 2–3 h and 2–3 e regions. (In **h**, the current switched between two stable positions for positive bias, with corresponding noise in dI/dV .)

varies from $5\text{ G}\Omega$ in the first Coulomb peak to $400\text{ k}\Omega$ at large negative V_G .

The observation of electron–hole symmetry poses severe restrictions on the QD system: the effective masses for holes and electrons should be equal and the QD should be free of disorder. Scattering by negatively charged impurities, for example, is repulsive for electrons but attractive for holes, so it would break electron–hole symmetry. A symmetric band structure has been predicted theoretically for graphite materials and carbon nanotubes³. In contrast, the absence of scattering has come as a surprise.

Figure 4 clarifies the correspondence between the electron and hole excitation spectra. On the right side of Fig. 4b, the situation for electrons is drawn (for $V_G > 0$) and on the left side for holes (for $V_G < 0$). The resulting excitations in transport characteristics as a function of V and V_G then lead to spectra as sketched in Fig. 4c, and as measured in Fig. 3.

A detailed analysis of the excitation spectrum requires calculations that are beyond the scope of this paper. The constant-interaction model provides the parameter range for more exact models. The change in orbital energy when adding a charge is given by $\Delta E \approx 4.3$ meV, independent of N . ΔE is the scale for the energy difference between single-particle states. Excitations of a smaller energy scale have to be related to interactions. The likely inter-

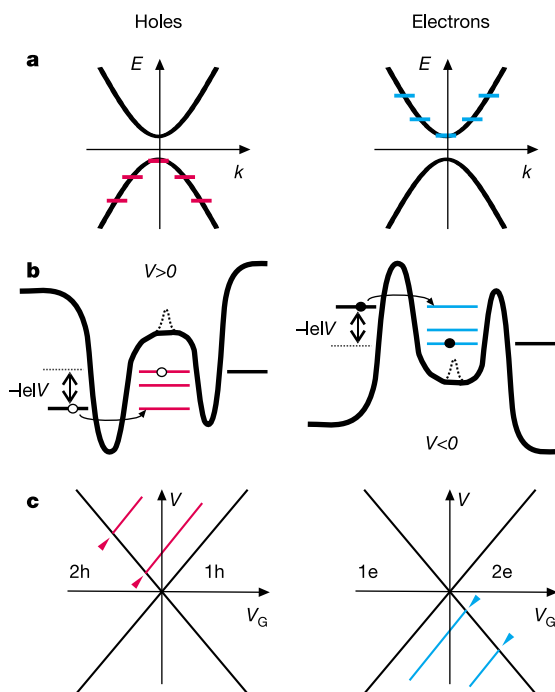


Figure 4 Electron–hole symmetry in semiconducting SWNTs. **a**, Band structure (energy E versus wave vector k) of a semiconducting nanotube illustrating symmetric valence and conduction bands. Owing to quantum confinement, the carriers occupy a set of discrete energy states, shown on the left for hole doping and on the right for electron doping. **b**, Energy diagrams showing transport of holes (left) and electrons (right) across a QD. The levels for the ground state and two excited states for $N = 2$ are drawn. The distance between equivalent levels on the right and left is the same, owing to equal effective electron and hole masses. The top hole level is accessed from the left for $V > 0$, whereas the top electron level aligns with the left lead Fermi energy for $V < 0$. The dotted line in the potentials shows the effect of a negatively charged scatterer, which breaks electron–hole symmetry. **c**, Excitation spectra resulting from the energy diagrams in **b**. New levels entering the bias window owing to excited states lead to lines in the dI/dV plots that run parallel to the diamond edges. (Note that, owing to asymmetric barriers, only excitation lines with positive slope become visible.) These diagrams can be compared to the experiment in Fig. 3.

actions in semiconducting nanotubes are: (1) exchange interaction between spins (for example, spin = 1 triplet states gain energy from the exchange interaction). Note that we observe an even-odd pattern, which seems to exclude ground states with spins $>1/2$. Excited states, however, can have spins $>1/2$. (2) Electron-phonon interactions. The vibrational modes in a suspended nanotube also have a discrete spectrum, which can show up in the excitation spectra¹⁹. Note that vibrational modes do not affect the addition energy spectrum of the ground states. (3) Electron-electron interactions. The value for the interaction strength parameter $U/\Delta E \approx 10$. Such a large $U/\Delta E$ ratio points to the presence of phenomena that are not captured by the constant-interaction model. Luttinger liquid models developed for finite length metallic nanotubes are not applicable to our few-particle nanotubes. A more appropriate starting point are the exact calculations for one-dimensional QDs. In the few-particle regime, the charge carriers tend to localize and maximize their separation, thereby forming a Wigner crystal²⁰. In such a Wigner state, the spectrum consists both of high-energy single-particle excitations and collective excitations at low energy²¹, similar to our experiment. Detailed calculations beyond the constant-interaction model and a comparison with the experimental results are necessary to establish the precise effect on transport from these interactions. □

Received 22 October 2003; accepted 13 April 2004; doi:10.1038/nature02568.

1. Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* (Saunders College, Orlando, 1976).
2. Dekker, C. Carbon nanotubes as molecular quantum wires. *Phys. Today* **52**, 22–28 (1999).
3. Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
4. Kouwenhoven, L. P., Austing, D. G. & Tarucha, S. Few-electron quantum dots. *Rep. Prog. Phys.* **64**, 701–736 (2001).
5. Tans, S. J. *et al.* Individual single-wall nanotubes as quantum wires. *Nature* **386**, 474–477 (1997).
6. Bockrath, M. *et al.* Single-electron transport in ropes of carbon nanotubes. *Science* **275**, 1922–1925 (1997).
7. Cobden, D. H. & Nygård, J. Shell filling in closed single-wall carbon nanotube quantum dots. *Phys. Rev. Lett.* **89**, 046803 (2002).
8. Ando, T. & Nakanishi, T. Impurity scattering in carbon nanotubes—Absence of back scattering. *Jpn. J. Appl. Phys.* **37**, 1704–1713 (1998).
9. McEuen, P. L., Bockrath, M., Cobden, D. H., Yoon, Y. & Louie, S. G. Disorder, pseudospins, and backscattering in carbon nanotubes. *Phys. Rev. Lett.* **83**, 5098–5101 (1999).
10. Bronikowski, M. J., Willis, P. A., Colbert, D. T., Smith, K. A. & Smalley, R. E. Gas-phase production of carbon single-walled nanotubes from carbon monoxide via the HiPco process: A parametric study. *J. Vacuum Sci. Technol. A* **19**, 1800–1805 (2001).
11. Fuhrer, M. S. *et al.* Crossed nanotube junctions. *Science* **288**, 494–497 (2000).
12. Nygård, J. & Cobden, D. H. Quantum dots in suspended single-wall carbon nanotubes. *Appl. Phys. Lett.* **79**, 4216–4218 (2001).
13. Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. Logic circuits with carbon nanotube transistors. *Science* **294**, 1317–1320 (2001).
14. Park, J. & McEuen, P. L. Formation of a p -type quantum dot at the end of an n -type carbon nanotube. *Appl. Phys. Lett.* **79**, 1363–1365 (2001).
15. Postma, H. W. C., Sellmeijer, A. & Dekker, C. Manipulation and imaging of individual single-wall carbon nanotubes with an atomic force microscope. *Adv. Mater.* **12**, 1299–1302 (2000).
16. Postma, H. W. C., Yao, Z. & Dekker, C. Electron addition and excitation spectra of individual single-wall carbon nanotubes. *J. Low-Temp. Phys.* **118**, 495–507 (2000).
17. Heinze, S. *et al.* Carbon nanotubes as Schottky barrier transistors. *Phys. Rev. Lett.* **89**, 106801 (2002).
18. Grabert, H. & Devoret, M. H. *Single Charge Tunneling* (Plenum, New York, 1992).
19. Park, H. *et al.* Nano-mechanical oscillations in a single- C_{60} transistor. *Nature* **407**, 57–60 (2000).
20. Wigner, E. On the interaction of electrons in metals. *Phys. Rev.* **46**, 1002–1011 (1934).
21. Häussler, W. & Kramer, B. Interacting electrons in a one-dimensional quantum dot. *Phys. Rev. B* **47**, 16353–16357 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. E. Smalley and co-workers for providing the high-quality HiPco nanotubes, and S. De Franceschi, J. Kong, K. Williams, Y. Nazarov, H. Postma, S. Lemay and J. Fernández-Rossier for discussions. We acknowledge the technical assistance of R. Schouten, B. van der Enden and M. van Oossanen. Financial support was obtained from the Dutch Organization for Fundamental Research (FOM).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.J.-H. (Pablo@qt.tn.tudelft.nl).

Electric polarization reversal and memory in a multiferroic material induced by magnetic fields

N. Hur, S. Park, P. A. Sharma, J. S. Ahn*, S. Guha & S.-W. Cheong

Department of Physics & Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA

* Present address: Department of Physics, Pusan National University, Pusan 609-735, Korea

Ferroelectric and magnetic materials are a time-honoured subject of study and have led to some of the most important technological advances to date. Magnetism and ferroelectricity are involved with local spins and off-centre structural distortions, respectively. These two seemingly unrelated phenomena can coexist in certain unusual materials, termed multiferroics^{1–11}. Despite the possible coexistence of ferroelectricity and magnetism, a pronounced interplay between these properties has rarely been observed^{6,12}. This has prevented the realization of multiferroic devices offering such functionality¹³. Here, we report a striking interplay between ferroelectricity and magnetism in the multiferroic TbMn_2O_5 , demonstrated by a highly reproducible electric polarization reversal and permanent polarization imprint that are both actuated by an applied magnetic field. Our results point to new device applications such as magnetically recorded ferroelectric memory.

Multiferroics, sometimes called magnetoelectrics, possess two or more switchable states such as polarization, magnetization or strain^{1,2}. Although there are a number of materials that possess both ferroelectricity and magnetism, it may be surprising that there need not necessarily be a large coupling between them. The non-trivial spin-lattice coupling in these multiferroics has been manifested through various forms, such as linear and bilinear magnetoelectric effects^{14,15}, polarization change through field-induced phase transition^{16,17}, magneto-dielectric effect^{5,7}, and dielectric anomalies at magnetic transition temperatures^{8,9}. Arguably, magnetoelectric effects have been studied the most in multiferroics such as perovskite-type BiFeO_3 (refs 3, 4) or BiMnO_3 (ref. 5), the boracite family^{6,7}, and the families of BaMF_4 (M, divalent transition metal ion)⁸, hexagonal RMnO_3 (R, rare earths)^{9,10}, and the rare-earth molybdates¹¹. In general, the applied magnetic field results in a small modulation in the spontaneous polarization. However, the complete (non-reversible) rotation of ferroelectric domains by magnetic fields has rarely been observed^{6,11}. Why and under what circumstances a large coupling should come about is a major open question, but this problem has proved difficult to tackle owing to the lack of materials that show such large coupling.

Here we report that the multiferroic TbMn_2O_5 exhibits a profound interplay between electrical polarization and the applied magnetic field. This is demonstrated by two discoveries: the highly reversible switching of electrical polarization using relatively low magnetic fields of 0–2 T; and the combined application of electric and magnetic fields, which leaves a permanent imprint in the polarization. This ‘reversible’ polarization switching and memory effect are distinct from the ‘non-reversible’ rotation of ferroelectric domains by magnetic fields^{6,11}.

Very little is actually known about the ferroelectricity and magnetism of TbMn_2O_5 , primarily owing to its chemical and structural complexity^{18–20}. The structure at room temperature is orthorhombic (space group, $Pbam$). The Mn^{4+} ions are octahedrally coordinated by oxygen, whereas Mn^{3+} ions are at the base centre of a square pyramid. Neutron diffraction studies have suggested a helical distribution of Mn moments rotating in the ab

plane at ~ 25 K. Below ~ 4.2 K, the Mn and Tb moments are believed to be amplitude-modulated at fixed angles to the ab plane^{21,22}. In addition, the Tb ordering transition was indirectly suggested through alternating-current magnetoelectric measurements¹⁸.

A delicate balance between various exchange interactions among the Tb^{3+} , Mn^{3+} and Mn^{4+} spins and the lattice polarization induces a number of phase transitions in TbMn_2O_5 . Figure 1a shows the temperature (T) dependence of the magnetic susceptibility, $\chi = M/H$ as well as the magnetic field (H) dependence of the magnetization of a TbMn_2O_5 crystal along the three principal crystallographic directions. (Single crystals of TbMn_2O_5 were grown using $\text{B}_2\text{O}_3/\text{PbO}/\text{PbF}_2$ flux in a Pt crucible. The flux was held at $1,280^\circ\text{C}$ for 15 hours and slowly cooled down to 950°C at a rate of 1°C per hour. Crystals grew in the form of black platelets as well as cubes with a typical size of $\sim 10\text{ mm}^3$.) A large magnetic anisotropy, primarily originating from the large spin-orbit coupling of the Tb moment, was clearly observed. $M(H)$ curves at 2 K in the inset of Fig. 1a show that the magnetic 'easy' direction is along the a axis, that the magnetization saturates for $H > 2T$ applied along the a axis, and that the corresponding saturation moment is $8.2\mu_B$ per formula unit, consistent with earlier studies¹⁸. Several distinct features occur in $\chi(T)$ at ~ 24 K, ~ 38 K and ~ 43 K, more clearly displayed in Fig. 1b. All three crystallographic directions show this behaviour, although it is more apparent along the b and c axes. Long-range antiferromagnetic order of the $\text{Mn}^{3+}/\text{Mn}^{4+}$ spins is known to occur at about 40 K. The 24 K feature has been identified

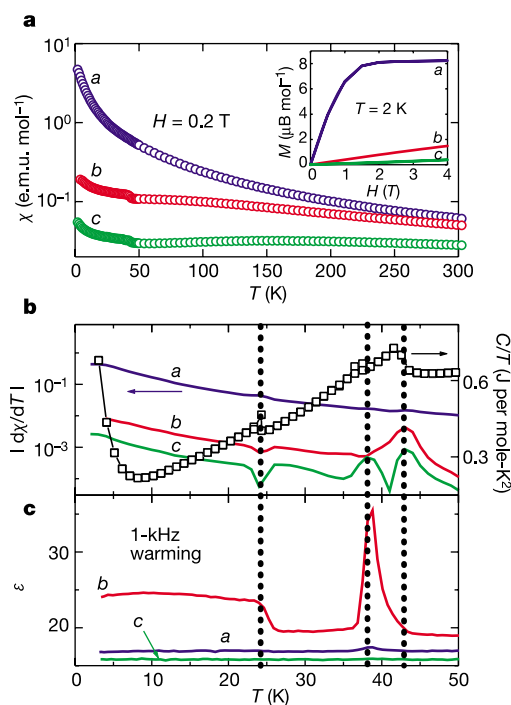


Figure 1 Coupled dielectric and magnetic properties at phase transitions. **a**, Temperature dependence of magnetic susceptibility of a TbMn_2O_5 crystal along three crystallographic directions, measured in 2 kOe. The inset shows magnetization versus magnetic field up to 4 T. **b**, The temperature derivative of the magnetic susceptibility along the three crystallographic directions as well as the temperature dependence of specific heat exhibit various low-temperature features, signalling a number of phase transitions. **c**, Dielectric constant versus temperature. The features of the dielectric constant along the b axis closely match those in Fig. 1b. For the dielectric measurement at 1 kHz, we used thin rectangular specimens with a cross-sectional area of $1\text{--}2\text{ mm}^2$ and a thickness of $0.2\text{--}0.5\text{ mm}$. The dielectric constants along the a and c directions show negligible anomaly. Certainly, this material is highly anisotropic in magnetic and dielectric properties.

as a Mn spin reorientation¹⁹. The structural complexity of this material allows for a number of magnetic exchange paths of comparable magnitude that can cause such a transition. Although there is no clear evidence in our $\chi(T)$ data, Tb magnetic ordering is suspected to occur below ~ 10 K (refs 18, 21, 22).

The specific heat, C (Fig. 1b), and dielectric constant, ϵ (Fig. 1c), also exhibit features that closely match those in χ . The large peak in ϵ along the b axis at 38 K signals the onset of ferroelectricity, just below the magnetic-ordering temperature. Earlier studies have shown a much-less-pronounced peak and a lower ferroelectric transition T (T_C), probably resulting from lower crystal quality¹⁸. There are no clear dielectric signatures along the a and c directions, indicating the dominant orientation of polarization, P , along the b axis. Intriguingly, the rise of the b -axis ϵ starts upon $\text{Mn}^{3+}/\text{Mn}^{4+}$ spin ordering at 43 K, suggesting a close relationship to the subsequent appearance of ferroelectricity along this direction. It has been suggested that the long-range magnetic ordering of Mn^{3+}

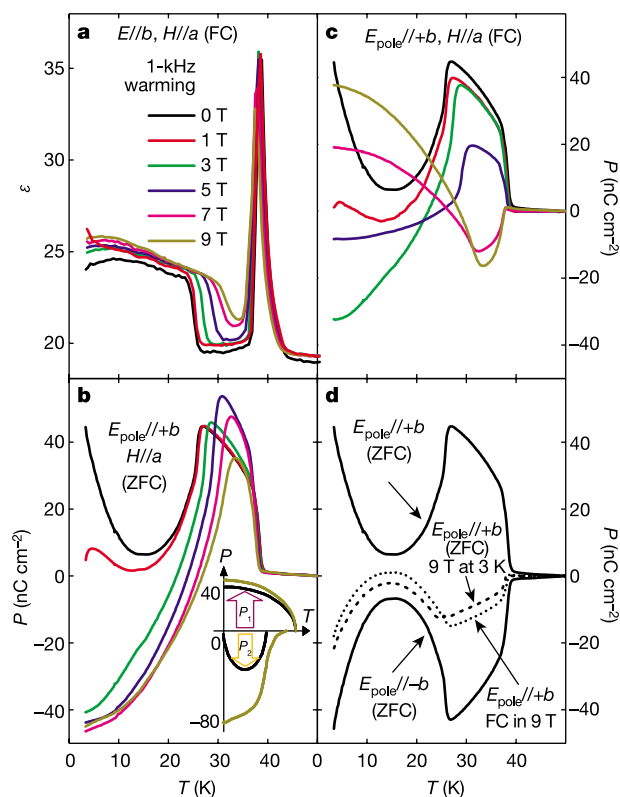


Figure 2 Magnetic-field dependence of ferroelectricity and memory effect.

a, Temperature dependence of 1-kHz dielectric constants along the b axis in various magnetic fields applied along the a axis. For clarity, only warming data are displayed. **b**, Temperature dependence of the total electric polarization along the b axis in magnetic fields. The polarization was calculated by integrating the measured pyroelectric current²³. Before each measurement, the sample was cooled from 120 K to 3 K in a static electric field, $E_{\text{pole}} = 4\text{ kV cm}^{-1}$ and zero magnetic field (ZFC); the pyroelectric current was measured while heating the sample at a rate of 4 K min^{-1} . Inset, the measured polarization may be the sum of a positive component (P_1) and a negative component (P_2). The inset shows a diagram of the proposed temperature dependence of P_1 and P_2 in $H = 0\text{ T}$ and 9 T . **c**, Total electric polarization in various magnetic fields, measured after cooling in a magnetic field (FC) along the a axis as well as an electric field E_{pole} along the $+b$ axis. **d**, Total electric polarization measured in zero magnetic field after cooling under four different conditions: E_{pole} along the $+b$ axis without magnetic field; E_{pole} along the $-b$ axis without magnetic field; E_{pole} along the $+b$ axis with a magnetic field of 9 T; and E_{pole} along the $+b$ axis without magnetic field followed by an application of 9 T at 3 K, which is again followed by removing E_{pole} before removing the magnetic field of 9 T. The applied magnetic field leaves a permanent imprint in the polarization.

and Mn^{4+} induces the ferroelectric transition through an additional distortion of the Jahn–Teller-distorted Mn^{3+} neighbours¹⁹. A pronounced ‘step’ in the $\epsilon(T)$ at 24 K appears simultaneously with respect to the $\text{Mn}^{3+}/\text{Mn}^{4+}$ spin reorientation. Although a detailed study of the magnetic ordering and accompanying polar atomic displacements has not yet been performed, these dielectric features at the magnetic transition temperatures signal non-trivial spin–lattice coupling. Note that the significant C rise below 10 K on cooling may suggest the slow onset of Tb magnetic order.

We attempted to observe how the dielectric behaviour evolves with H along the a axis; H induces a drastic increase of magnetization at low temperatures. The magnetic field had a strong influence on ϵ along the b axis, whereas little change was observed along the a and c axes, as depicted in Fig. 2a. As evident in the figure, the 38-K ferroelectric transition gradually decreases in T with increasing H , whereas the ‘step’ anomaly dramatically increases in T with increasing H .

To observe the influence of H on P directly, we measured the temperature-dependent pyroelectric current and magnetic-field-dependent magnetoelectric current along the b axis. Magnetic field was always applied along the a axis; curiously, only fields along this direction appeared to have any effect at all. Figure 2b and c

displays the total $P(T)$, the sum of the spontaneous P and the H -induced P , in various magnetic fields. All measurements were performed by poling the sample along the b axis in a static electric field, $E_{\text{pole}} = 4 \text{ kV cm}^{-1}$. At $H = 0$, P starts to increase at 38 K, confirming the onset of ferroelectricity signalled by the peak in $\epsilon(T)$. The lower-temperature ‘step’ in $\epsilon(T)$ correlates with a sudden decrease in P on cooling, both of which behave systematically with increasing H . In general, an anomaly in $\epsilon(T, H)$ indicates a sudden change of P (for example, rotation or onset). As is evident in Fig. 2b, even modest fields as low as 1 T had a surprisingly profound effect on P . The direction of the P at low T could be completely switched with the application of only 2 T. This field effect can be seen more dramatically by ‘poling’ the sample with magnetic field cooling (FC) in addition to applying E_{pole} . As shown in Fig. 2c, the FC polarization at low temperatures initially decreases and becomes negative with increasing H , but for $H > 3 \text{ T}$, it increases again, and eventually becomes positive.

This seemingly quite complicated temperature and magnetic-field dependence may be explained using the concept of ferrielectricity^{23–26}. The net polarization can be considered to be the sum of a few different components, each of which may have a different T and H dependence. It is natural to consider such a model because of the structural complexity inherent in this material. From the shape of the polarization data for $H = 0$, we conjecture that the total P is composed of a positive component (P_1) which spontaneously appears at $T_C \approx 38 \text{ K}$ and a negative component (P_2) at $T_C \approx 24 \text{ K}$ as shown in the inset of Fig. 2b. We emphasize that the absence of the a - and c -axes $\epsilon(T)$ anomaly in magnetic fields indicates that P_1 and P_2 always lie along the $\pm b$ axis. Because the direction of P can be switched completely by an oppositely directed E_{pole} , as shown in Fig. 2d, we can assume that the direction of P_2 is always opposite to that of P_1 . Once E_{pole} and H fix the direction of P_1 , P_2 always chooses the opposite direction. Judging from the overall behaviour of the polarization in Fig. 2b, P_1 does not appear to be very sensitive to H , while P_2 changes considerably, particularly below $\sim 15 \text{ K}$.

There is clearly a distinct difference between FC poling and poling in zero magnetic field (ZFC), as shown by comparing Fig. 2b and c, but this combined magnetic- and electric-field poling process can also affect P even after removing H , as illustrated in Fig. 2d, for which all measurements were taken in zero field. In this figure, applying E_{pole} along the $+b$ direction produces a net positive total P , and reversing E_{pole} to be along the $-b$ direction naturally produces a net negative total P . However, poling with E_{pole} along the $+b$ direction and H along the a direction surprisingly results in a net negative total P measured at $H = 0$. A net negative total P was also obtained when a high magnetic field was applied at 3 K after ZFC poling along the $+b$ direction, but E_{pole} was removed before removing H . This seemingly contradictory result can be understood with the scenario we have considered. At low enough T ($< \sim 20 \text{ K}$), the negative P_2 state should become dominant in high applied magnetic fields, as evident in Fig. 2b. The simultaneous application of a high H and a positive E_{pole} will align a large P_2 along the $+b$ axis and consequently, P_1 along the $-b$ axis (P_1 and P_2 always remain in opposite directions). After removing E_{pole} and H , the P_1 state should now become dominant, but P_1 is in the direction opposite to E_{pole} , consistent with the above experimental results. This phenomenon is truly unique in the sense that a magnetic field can leave a permanent polarization imprint even after removing H . Thus, the polarization state ‘remembers’ the applied electric as well as magnetic field.

The aforementioned polarization switching by applied magnetic field results in a giant magnetocapacitance effect^{3,7}. Figure 3a shows $\epsilon(H)$ at 3 and 28 K. The dielectric constant exhibits a maximum change of 13% at 3 K and 20% at 28 K. At 3 K, a broad peak occurs in ϵ at around 1 T, with a small amount of hysteresis, corroborating the behaviour of $P(H)$ in Fig. 3b. This peak rapidly smears out as T

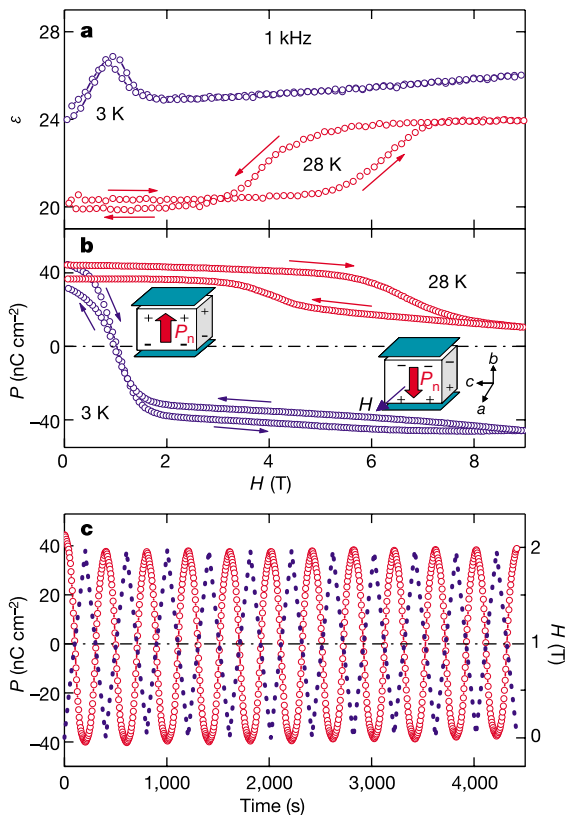


Figure 3 Reproducible polarization reversal by magnetic fields. **a**, Dielectric constant versus applied magnetic field at 3 and 28 K. **b**, Change of total electric polarization by applied magnetic fields at 3 and 28 K. The magnetic-field dependence of the total electric polarization was obtained by measuring the magnetoelectric current as a function of magnetic field, which was varied linearly with time at the uniform rate of 100 Oe s^{-1} . The magnetoelectric current was measured after cooling with E_{pole} along the $+b$ axis without magnetic field. The total polarization was obtained by adding the spontaneous polarization to the field-induced polarization, which was calculated from the magnetoelectric current. The cartoon shows the orientation of the net polarization ($P_n = P_1 + P_2$) in zero field and high fields $> 2 \text{ T}$. **c**, Polarization flipping at 3 K by linearly varying magnetic field from 0 to 2 T. These results clearly display highly reproducible polarization switching by magnetic fields.

increases and diminishes at ~ 10 K, suggesting a relation to Tb magnetic ordering. At 28 K, ϵ exhibits a field-induced 'step' anomaly with a larger hysteresis rather than a peak, consistent with the H dependence of the step feature in $\epsilon(T)$ shown in Fig. 2a as well as $P(H)$ in Fig. 3b.

The electric polarization shows a startling magnetic-field-induced reversal at low T , shown in Fig. 3b. At 3 K, the direction of P is completely reversed with a rather low H , corresponding to a change in P of ~ 80 nC cm $^{-2}$ at 2 T. This change in P for H below 2 T is considerably large and nonlinear, suggesting that this behaviour stems from H -induced phase transitions^{16,17}, possibly associated with the H -induced change of Tb magnetism. However, the P change for high H is rather linear with H , indicating that the linear magnetoelectric effect^{14,27} becomes active above 2 T. These results are certainly consistent with the behaviour of magnetization, which saturates for H along the a axis above 2 T. Note that the magnetic symmetry $P_{2a}m'c'2_1$ (magnetic point group, $m'm'2$) in zero field proposed from a neutron diffraction study^{18,22} allows only non-zero diagonal elements of the linear magnetoelectric tensor^{14,15}, but H along the a axis above 2 T appears to induce ferromagnetic Tb ordering, changing the magnetic symmetry and thus allowing a large off-diagonal linear magnetoelectric coefficient, α_{21} , of about -21 ps m $^{-1}$. We also point out that the reversible P flip along the $\pm b$ axis with $H = 0-2$ T should be distinguished from the non-reversible 180° switching of the spontaneous polarization induced by the 90° rotation of magnetic field in $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$ (ref. 6), the non-reversible 180° switching of the spontaneous polarization of a portion of the ferroelectric domains in $\text{Tb}_2(\text{MoO}_4)_3$ (ref. 11), and the 90° rotation of P by applied $H \approx 5$ T in TbMnO_3 (ref. 12).

We performed a sequential flipping of P by slowly varying H linearly from 0 to 2 T. A strikingly reversible and reproducible variation of the polarization was observed without any noticeable decay in its magnitude, as displayed in Fig. 3c. By combining this highly reproducible P reversal with the ability to leave a permanent 'imprint' in the polarization with an applied magnetic field demonstrated in Fig. 3c and Fig. 2d, respectively, one can envision a ferroelectric memory that can be magnetically recorded. □

Received 5 November 2003; accepted 19 April 2004; doi:10.1038/nature02572.

- Smolenski, G. A. & Chupis, I. E. Ferroelectromagnets. *Usp. Fiz. Nauk.* **137**, 415–448 (1982); *Sov. Phys. Usp.* **25**, 475–493 (1982).
- Schmid, H. Multi-ferroic magnetoelectrics. *Ferroelectrics* **162**, 317–338 (1994).
- Wang, J. *et al.* Epitaxial BiFeO_3 multiferroic thin film heterostructures. *Science* **299**, 1719–1722 (2003).
- Smolenskii, G. A., Yudin, V. M., Sher, E. S. & Stolypin, Yu. E. Antiferromagnetic properties of some perovskites. *Sov. Phys. JETP* **16**, 622–624 (1963).
- Kimura, T. *et al.* Magnetocapacitance effect in multiferroic BiMnO_3 . *Phys. Rev. B* **67**, 180401 (2003).
- Ascher, E., Rieder, H., Schmid, H. & Stössel, H. Some properties of ferromagnetoelectric nickel-iodine borate, $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$. *J. Appl. Phys.* **37**, 1404–1405 (1966).
- Baturov, L. N., Al'shin, B. I. & Yarmukhamedov, Yu. N. Nonlinear magnetoelectric and dielectric properties of Ni-I borate. *Fiz. Tverd. Tela* **20**, 2254–2259 (1978); *Sov. Phys. Solid State* **20**, 1300–1303 (1978).
- Scott, J. F. Phase transitions in BaMnF_4 . *Rep. Prog. Phys.* **12**, 1056–1084 (1979).
- Katsufuji, T. *et al.* Dielectric and magnetic anomalies and spin frustration in hexagonal RMnO_3 ($R = \text{Y, Yb, and Lu}$). *Phys. Rev. B* **64**, 104419 (2001).
- Fiebig, M., Lottermoser, Th., Fröhlich, D., Goltsev, A. V. & Pisarev, R. V. Observation of coupled magnetic and electric domains. *Nature* **419**, 818–820 (2002).
- Ponomarev, B. K., Ivanov, S. A., Popov, Yu. F., Negrii, V. D. & Red'kin, B. S. Magnetoelectric properties of some rare earth molybdates. *Ferroelectrics* **161**, 43–48 (1994).
- Kimura, T. *et al.* Magnetic control of ferroelectric polarization. *Nature* **426**, 55–58 (2003).
- Hill, N. A. Why are there so few magnetic ferroelectrics? *J. Phys. Chem. B* **104**, 6694–6709 (2000).
- Rivera, J.-P. On definitions, units, measurements, tensor forms of the linear magnetoelectric effect and on a new dynamic method applied to Cr-Cl borate. *Ferroelectrics* **161**, 165–180 (1994).
- Schmid, H. On a magnetoelectric classification of materials. *Int. J. Magn.* **4**, 337–361 (1973).
- Popov, Yu. F. *et al.* Low-temperature phase transition in EuMn_2O_5 . *Physica B* **284–288**, 1402–1403 (2000).
- Popov, Yu. F. *et al.* Features of the magnetoelectric properties of BiFeO_3 in high magnetic fields. *J. Low-Temp. Phys.* **27**, 478–479 (2001).
- Saito, K. & Kohn, K. Magnetoelectric effect and low-temperature phase transitions of TbMn_2O_5 . *J. Phys. Condens. Matter* **7**, 2855–2863 (1995).
- Golovenchits, E. I., Morozov, N. V., Sanina, V. A. & Sapozhnikova, L. M. Correlation of magnetic and dielectric properties in EuMn_2O_5 single crystals. *Sov. Phys. Solid State* **34**, 56–59 (1992).
- Goshen, S., Mukamel, D., Shaked, H. & Shtrikman, S. Magnetic symmetry and ferroelectricity induced by antiferromagnetic transitions. *J. Appl. Phys.* **40**, 1590–1592 (1969).

- Buisson, G. Structures magnétiques sinusoïdales et hélicoïdales de la terre rare dans TMn_2O_5 . *Phys. Status Solidi A* **17**, 191–198 (1973).
- Gardner, P. P., Wilkinson, C., Forsyth, J. B. & Wanklyn, B. M. The magnetic structures of the rare-earth manganates ErMn_2O_5 and TbMn_2O_5 . *J. Phys. C* **21**, 5653–5661 (1988).
- Inomata, A. & Kohn, K. Pyroelectric effect and possible ferroelectric transition of helimagnetic GdMn_2O_5 , TbMn_2O_5 and YMn_2O_5 . *J. Phys. Condens. Matter* **8**, 2673–2678 (1996).
- Sawada, A., Ohya, S., Ishibashi, Y. & Takagi, Y. Ferroelectric phase transition in $(\text{NH}_4)_2\text{SO}_4\cdot\text{K}_2\text{SO}_4$ mixed crystals. *J. Phys. Soc. Jpn* **38**, 1408–1414 (1975).
- Sawada, A. & Tomatsu, M. Ferrielectric behaviors in $\text{Li}_2\text{Ge}_2\text{O}_{15}$ crystals. *Ferroelectrics* **137**, 297–302 (1992).
- Maisonneuve, V., Cajipe, V. B., Simon, A., Von Der Muhll, R. & Ravez, J. Ferrielectric ordering in lamellar CuInP_2S_6 . *Phys. Rev. B* **56**, 10860–10868 (1997).
- Nakamura, N. & Kohn, K. Magnetoelectric effect of rare earth manganese RMn_2O_5 . *Ferroelectrics* **204**, 107–114 (1997).

Acknowledgements This work was supported by the National Science Foundation-MRSEC, and J.S.A. was primarily supported by the Korea Science and Engineering Foundation through the Center for Strongly Correlated Materials Research, Seoul National University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.H. (namjung@physics.rutgers.edu).

Evidence from massive siderite beds for a CO₂-rich atmosphere before ~ 1.8 billion years ago

Hiroshi Ohmoto¹, Yumiko Watanabe¹ & Kazumasa Kumazawa²

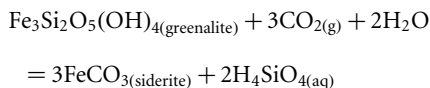
¹Astrobiology Research Center of the NASA Astrobiology Institute and Department of Geosciences, The Pennsylvania State University, University Park, PA 16802, USA

²Oyo Corporation, Miyazaki Branch, Oshima-cho, Miyazaki City, 0995-61, Japan

It is generally thought that, in order to compensate for lower solar flux and maintain liquid oceans on the early Earth, methane must have been an important greenhouse gas before ~ 2.2 billion years (Gyr) ago^{1–5}. This is based upon a simple thermodynamic calculation that relates the absence of siderite (FeCO_3) in some pre-2.2-Gyr palaeosols to atmospheric CO_2 concentrations that would have been too low to have provided the necessary greenhouse effect¹. Using multi-dimensional thermodynamic analyses and geological evidence, we show here that the absence of siderite in palaeosols does not constrain atmospheric CO_2 concentrations. Siderite is absent in many palaeosols (both pre- and post-2.2-Gyr in age) because the O_2 concentrations and pH conditions in well-aerated soils have favoured the formation of ferric (Fe^{3+})-rich minerals, such as goethite, rather than siderite. Siderite, however, has formed throughout geological history in subsurface environments, such as euxinic seas, where anaerobic organisms created H_2 -rich conditions. The abundance of large, massive siderite-rich beds in pre-1.8-Gyr sedimentary sequences and their carbon isotope ratios indicate that the atmospheric CO_2 concentration was more than 100 times greater than today, causing the rain and ocean waters to be more acidic than today. We therefore conclude that CO_2 alone (without a significant contribution from methane) could have provided the necessary greenhouse effect to maintain liquid oceans on the early Earth.

The current popular model of a methane-rich Archaean atmosphere was first proposed by Rye *et al.*¹, who suggested that pre-2.2-Gyr palaeosols were characterized by the absence of iron-bearing minerals in the upper parts and by the presence of chlorite (a Fe^{2+} - and Fe^{3+} -rich aluminous silicate) in the lower parts of soil sections. They assumed that the precursor of chlorite before metamorphism

was greenalite (a Fe^{2+} -silicate) and that the Fe^{2+} leached by rain water from the upper soil section precipitated greenalite, not siderite (a Fe^{2+} -carbonate), in the lower soil section. Rye *et al.*¹ evaluated the stability relation of greenalite and siderite from the following reaction:



The stability relation of the two minerals at a given temperature and pressure condition depends on only two parameters: the partial pressure of gaseous CO_2 and the concentration of aqueous silica in solution. Using thermodynamic data from the literature, Rye *et al.*¹ calculated that greenalite, not siderite, becomes stable at $p_{\text{CO}_2} < 10^{-2.2}$ atm at 10°C (and $< 10^{-1.7}$ atm at 30°C) when $\text{H}_4\text{SiO}_4(\text{aq})$ is between $\sim 10^{-4}$ and $\sim 10^{-3.4} m$ (where m is concentration in moles per kg H_2O) (that is, ~ 6 to ~ 24 p.p.m. $\text{SiO}_2(\text{aq})$, a typical range in groundwater). By further assuming soil p_{CO_2} values were between $\sim 60\%$ and 100% of the atmospheric p_{CO_2} , they concluded that the maximum p_{CO_2} in the pre-2.2-Gyr atmosphere was $\sim 10^{-1.4}$ atm.

This maximum p_{CO_2} value is, however, less than the 'best guess' p_{CO_2} values (for example, $\sim 10^{-0.7}$ atm at 2.75 Gyr) estimated by Kasting⁶ from a one-dimensional climatic model where the primary greenhouse gas was CO_2 (curve $\text{CO}_2(\text{K})$ in Fig. 1). Because of this p_{CO_2} discrepancy, Rye *et al.*¹ suggested that another greenhouse gas, most probably methane, was important in the pre-2.2-Gyr atmosphere. Subsequently, a revised climatic model^{4,5} suggests values of $\sim 1,000$ p.p.m. ($\sim 10^{-3}$ atm) CH_4 and $\sim 2,500$ p.p.m. ($\sim 10^{-2.6}$ atm) CO_2 for the atmosphere at 2.8 Gyr ago (curves $\text{CO}_2(\text{P})$ and $\text{CH}_4(\text{P})$ in Fig. 1). For comparison, today's atmosphere contains only ~ 1 p.p.m. ($\sim 10^{-6}$ atm) CH_4 and ~ 350 p.p.m. ($\sim 10^{-3.5}$ atm) CO_2 (Fig. 1). Here we show a serious flaw in the thermodynamic approach by Rye *et al.*¹, which led to the suggestion of a methane-rich Archaean atmosphere²⁻⁵.

Rye *et al.*¹ neglected two important facts: (1) the formation of siderite (and any other Fe-bearing minerals) depends not only on p_{CO_2} and $m\text{H}_4\text{SiO}_4(\text{aq})$, but also on additional geochemical parameters, including p_{O_2} , pH and concentrations of Fe^{2+} species in solution; and (2) siderite is a very abundant mineral in pre-1.8-Gyr sedimentary rocks, especially in banded iron formations (BIFs). For example, Fig. 2a illustrates that the formation of a Fe^{2+} -rich mineral (for example, greenalite, minnesotaite, fayalite, annite or siderite) requires an extremely low p_{O_2} condition ($< 10^{-6.5 \pm 5}$ atm). Many researchers agree that the atmospheric p_{O_2} was much greater than $\sim 10^{-60}$ atm throughout geologic history, although suggested values vary greatly: for example, $\sim 10^{-13}$ atm (ref. 6), $\sim 10^{-4}$ atm (ref. 2) and ~ 0.2 atm (ref. 7) at ~ 2.75 Gyr ago (Fig. 2a). Soil p_{O_2} values are

typically less than the atmospheric p_{O_2} value owing to reactions with Fe^{2+} -bearing minerals and organic matter, but generally remain $\geq 10^{-60}$ atm in well-aerated soils⁸ (see Methods). At $p_{\text{O}_2} \geq 10^{-60}$ atm, siderite (or any ferrous-rich mineral) becomes unstable regardless of the p_{CO_2} value; instead, ferric-rich minerals, such as goethite (FeOOH), become stable (Fig. 2a).

Goethite transforms to haematite (Fe_2O_3) at temperatures $\geq 80^\circ\text{C}$ owing to dehydration⁸. A thick haematite-enriched zone, which was most probably a goethite-enriched zone before regional metamorphism, has been recognized in the upper parts of the ~ 2.3 -Gyr Ville Marie palaeosols⁹ in Canada, the ~ 2.3 -Gyr Hakkalampi palaeosols¹⁰ in Finland and the ~ 2.2 -Gyr Hekpoort palaeosols¹¹ in South Africa. The chlorite in a lower part of the Hekpoort palaeosols is a Fe^{3+} -rich aluminous silicate¹⁰, not pure Fe^{2+} silicate as assumed by Rye *et al.*¹. Furthermore, ferri-stilpnomelane (another Fe^{3+} -rich aluminous silicate) is abundant in the uppermost part of the 2.6-Gyr Schagen palaeosol¹². Many palaeosols have lost the uppermost parts through erosion during the deposition of overlying rock units, but the remaining soil profiles have retained the original characteristics of Fe^{3+} enrichments from the parental values⁷. (Although some earlier researchers wondered whether the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratios of Precambrian palaeosols were increased by modern weathering, recent investigators have concluded that many palaeosol sections have retained the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratios of original palaeosols^{2,7,10,11}). Therefore, both the thermodynamic predictions and natural examples clearly indicate that goethite and Fe^{3+} -rich aluminous silicates were stable minerals during soil formation both before and after 2.2 Gyr ago. The absence of siderite in the pre-2.2-Gyr soils that were investigated by Rye *et al.*¹ and in many younger palaeosols was not because the atmospheric p_{CO_2} was $< 10^{-1.4}$ atm as they suggested, but because the p_{O_2} values in well-aerated soils were typically much greater than $10^{-65 \pm 5}$ atm (Fig. 2a; Methods). In fact, Fe-rich carbonates formed in the ~ 2.6 -Gyr Schagen palaeosols¹² during periods when the soils were submerged in a shallow water body (pond or lake) where anaerobes created anoxic conditions (see Methods).

Another reason for the absence of siderite in well-aerated soils, both before and after 2.2 Gyr ago, is that siderite is only stable at pH conditions of more than ~ 1.5 units above the rainwater pH values (Fig. 2b). However, increases in soil pore-fluid pH (due to reactions with minerals) are typically by less than ~ 1.5 units (ref. 8).

Although Fe^{2+} -rich carbonates (siderite, ankerite and ferroan dolomite) are rare in well-aerated soils and palaeosols, they are not uncommon in sedimentary rocks (marine and freshwater) of all geologic age. For siderite formation, all of the following conditions must be satisfied (Fig. 2): (1) a very low p_{O_2} ($\leq 10^{-60}$ atm); (2) a very high p_{CO_2} ($> 10^{-1.4 \pm 0.2}$ atm); (3) a slightly acidic to near-neutral pH (~ 5.5 to ~ 7.5); and (4) a high concentration of ΣFe^{2+} ($= \text{Fe}^{2+} + \text{FeHCO}_3^-$) in water, most probably within a range of ~ 0.1 to ~ 100 p.p.m. ($m\Sigma\text{Fe}^{2+} = \sim 10^{-6}$ to $\sim 10^{-3}$), corresponding to the typical ΣFe^{2+} concentrations in low-temperature ($< 200^\circ\text{C}$) natural solutions⁸.

Archaean oceans may have been considerably warmer and much more silica-rich than today¹³ ($\text{SiO}_2(\text{aq}) = \sim 6$ p.p.m. today⁸). The p_{CO_2} value for the minnesotaite/siderite boundary, assuming a $\text{SiO}_2(\text{aq})$ concentration of 24 p.p.m., is $10^{-1.7}$, $10^{-1.6}$ and $10^{-1.3}$ atm at 10° , 25° and 50°C , respectively. However, if the $\text{SiO}_2(\text{aq})$ concentration of the Archaean oceans was as high as ~ 100 p.p.m., it becomes $10^{-1.2}$ atm at 25°C . Considering these variations, we propose here a p_{CO_2} value of $10^{-1.4 \pm 0.2}$ atm as the minimum p_{CO_2} requirement for siderite formation throughout geologic history. A combination of high p_{CO_2} and low pH creates carbonate-rich water with a ΣCO_3^{2-} ($\approx \text{CO}_2(\text{aq}) + \text{HCO}_3^-$) concentration between $\sim 10^{-2.5}$ and $\sim 10^{-0.5} m$ (compared with $\sim 10^{-3} m$ in today's open oceans⁸) (Fig. 2b). Subsurface waters (for example, lower water columns in euxinic seas; pore fluids in soils and sediments) may acquire all the necessary conditions for siderite

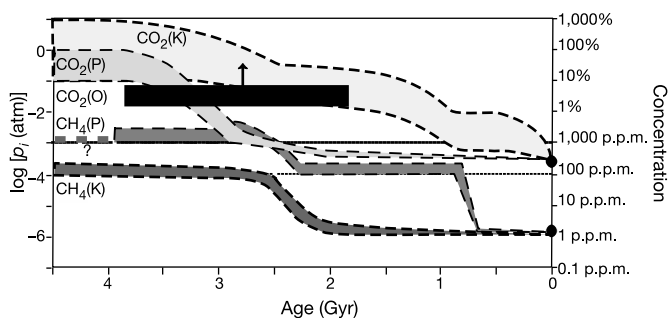


Figure 1 Proposed models for the evolution of atmospheric CO_2 and CH_4 . Curve $\text{CO}_2(\text{K})$ is reproduced from Kasting⁶; curve $\text{CH}_4(\text{K})$ was constructed from descriptions in Kasting⁶; curves $\text{CO}_2(\text{P})$ and $\text{CH}_4(\text{P})$ were constructed from the values estimated by Pavlov *et al.*^{4,5,27} at 3.9 Gyr, 2.8 Gyr and 2.3–0.75 Gyr ago; and line $\text{CO}_2(\text{O})$ is from this study. p_i refers to the partial pressure of a gaseous species i .

formation through a series of microbial and non-microbial reactions involving Fe-bearing minerals and organic matter (see Methods).

The major mechanism for siderite formation seems to have changed through geologic history, because distinct differences exist in the abundance, the occurrence and the chemical and isotopic compositions of siderite crystals between those younger and those older than ~ 1.8 Gyr. In post-1.8-Gyr sedimentary formations, Fe-rich carbonates rarely form massive beds but typically occur as disseminated crystals, cements and nodules in organic carbon-rich clastic sediments (mudstones and sandstones). The carbonate content is generally less than 10 vol.%, and the Fe/Ca ratios and $\delta^{13}\text{C}$ values are highly variable on the millimetre scale^{14,15}. These observations suggest that these siderite crystals formed during the diagenetic stage of host sediments, and that the necessary conditions for siderite formation were locally created for the pore fluids in organic carbon-rich sediments. In fact, the p_{O_2} - p_{CO_2} -pH conditions of subsurface waters with abundant anaerobes (for example, water-logged soils, euxinic lakes and seas) typically fall around the goethite-siderite boundary⁸ (Fig. 2a).

Many modern siderite crystals, especially those formed in sulphate-poor terrestrial environments (for example, ponds and lakes), possess very positive and variable $\delta^{13}\text{C}$ values (for example, $\sim +10\text{‰}$; Fig. 3a), suggesting the use of the ΣCO_3^{2-} generated from the decomposition of organic matter by methanogens^{14,15} (see reaction (6) in Methods). In contrast, siderite crystals that formed in sulphate-rich environments (for example, marine sediments) tend to possess negative and variable $\delta^{13}\text{C}$ values (Fig. 3b) owing to

the significant contribution of the ΣCO_3^{2-} generated from the decomposition of organic matter by non-methanogenic microbes, especially sulphate reducers¹⁵ (see reactions (1)–(5) in Methods).

In contrast to the siderite occurrences in young sediments, Fe-rich carbonates in many pre-1.8-Gyr sedimentary sequences formed massive beds of tens to hundreds of metres in thickness, and hundreds to thousands of thousands of square kilometres in aerial extent¹⁶. In fact, Fe-rich carbonates are more abundant than Fe-oxides (haematite and magnetite) in many BIFs that are mostly ~ 3.8 to ~ 1.8 Gyr in age¹⁶. Primarily on the basis of rare-earth-element characteristics, many investigators agree that the iron in BIFs was derived from oceanic basalts by submarine hydrothermal fluids^{16–18}. However, opinions vary as to whether the hydrothermal fluids primarily discharged on mid-ocean ridges creating global Fe^{2+} -rich oceans, or in BIF-forming basins creating local Fe^{2+} -rich brine pools, such as those in the Red Sea¹⁷. The widespread occurrence of massive siderite beds in pre-1.8-Gyr sedimentary sequences suggests that the required conditions for siderite formation were more common and widespread in pre-1.8-Gyr oceans. These conditions were most probably created by a CO_2 -rich atmosphere (see below).

We have carried out petrological, mineralogical and geochemical investigations of massive siderite beds in three major sedimentary sequences, the 1.9-Gyr Gunflint Iron Formation, the 2.6-Gyr Wittenoon Formation and the 2.75-Gyr Helen Iron Formation (Methods and Supplementary Tables). The results of our investigations support the following suggestions made by previous inves-

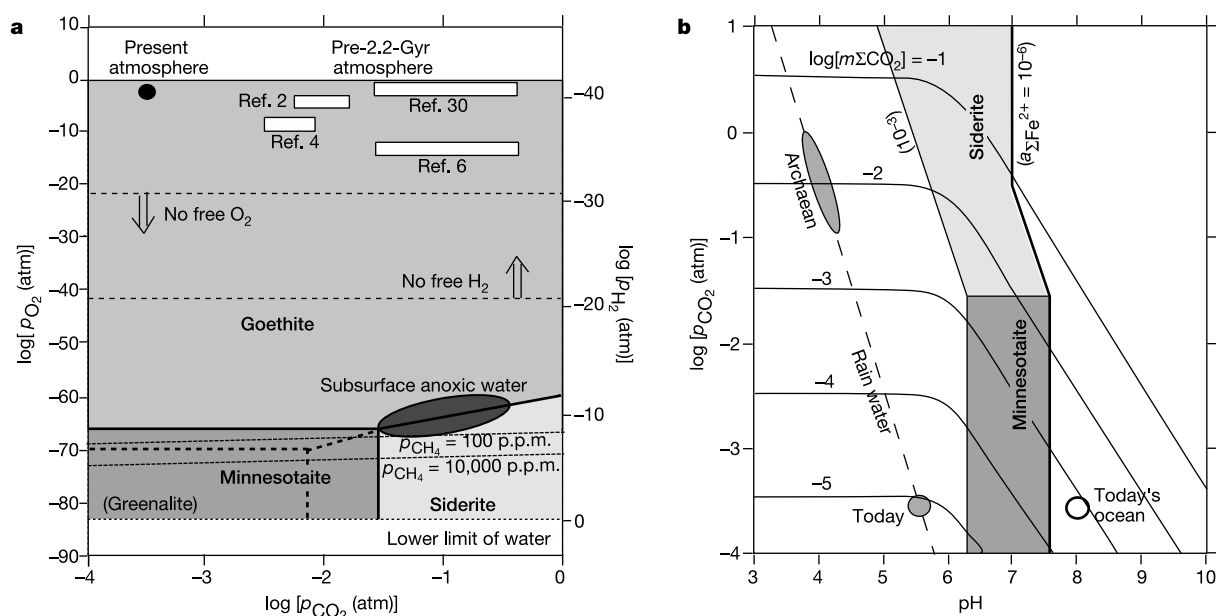


Figure 2 Thermodynamic conditions for the formation of siderite. **a**, A projection of the stability fields of major Fe-bearing minerals on a p_{O_2} (g)– p_{CO_2} (g) plane at $T = 25^\circ\text{C}$ and $\text{SiO}_{2(\text{aq})} = 10^{-3.4}$ moles per kg H_2O (24 p.p.m.); the diagram was constructed using the available thermodynamic data^{28,29}. Note that when $\text{SiO}_{2(\text{aq})} > 10^{-4.35}$ (>2.7 p.p.m.), the stability field of greenalite (shown by dotted lines) disappears but that of minnesotaite appears. The maximum p_{O_2} values for other Fe^{2+} -rich minerals, such as fayalite (olivine) and annite (biotite), fall within ± 5 log units of the minnesotaite–goethite boundary²⁹. Also shown are the p_{O_2} – p_{CO_2} values estimated by previous researchers^{2,4,6,7} for the ~ 2.75 -Gyr atmosphere. Although methane (CH_4) is thermodynamically unstable at $p_{\text{O}_2} \geq 10^{-60}$ atm, a significant amount of CH_4 may survive at higher p_{O_2} conditions, as in today's atmosphere, if the production rate of biogenic CH_4 exceeds the oxidation rate. The shaded area represents the typical conditions for modern subsurface anoxic waters with abundant anaerobic organisms⁸. **b**, A projection of the formational conditions of siderite and minnesotaite on a p_{CO_2} (g)–pH plane when $p_{\text{O}_2} \leq 10^{-60}$ atm,

$\text{SiO}_{2(\text{aq})} = 10^{-3.4}$ m and $\Sigma\text{Fe}^{2+} = 10^{-6}$ – 10^{-3} m. Henry's law constant relates $p_{\text{CO}_2}(\text{g})$ to the concentration of aqueous CO_2 in solution ($m\text{CO}_{2(\text{aq})}$), and the first and second dissociation constants of carbonic acids relate $m\text{CO}_{2(\text{aq})}$ to $m\text{HCO}_3^-$ and $m\text{CO}_3^{2-}$ at a specific pH. $m\Sigma\text{CO}_3^{2-} = m\text{CO}_2(\text{aq}) + m\text{HCO}_3^- + m\text{CO}_3^{2-}$. At $T = 25^\circ\text{C}$, rainwater pH is related to atmospheric p_{CO_2} by: $\text{pH} = 3.91 - 0.5 \log p_{\text{CO}_2}(\text{g})$ (ref. 30), suggesting that the Archean rainwater pH was between ~ 4.4 and ~ 3.9 , if the atmospheric p_{CO_2} was between 0.1 and 1.0 atm. Typical pH values of modern surface waters on land (soil-, ground- and river waters) are ~ 7 and those of ocean waters are ~ 8.1 (ref. 8); the increases from the rainwater pH (5.7) were caused by reactions with various minerals and evaporation. As a first-order approximation, we may expect that the differences in pH values among Archean surface waters were similar to those of modern surface waters, suggesting that the pH values of normal Archean oceans were between ~ 5.5 and ~ 7.5 . This pH range coincides with the formational conditions for siderite.

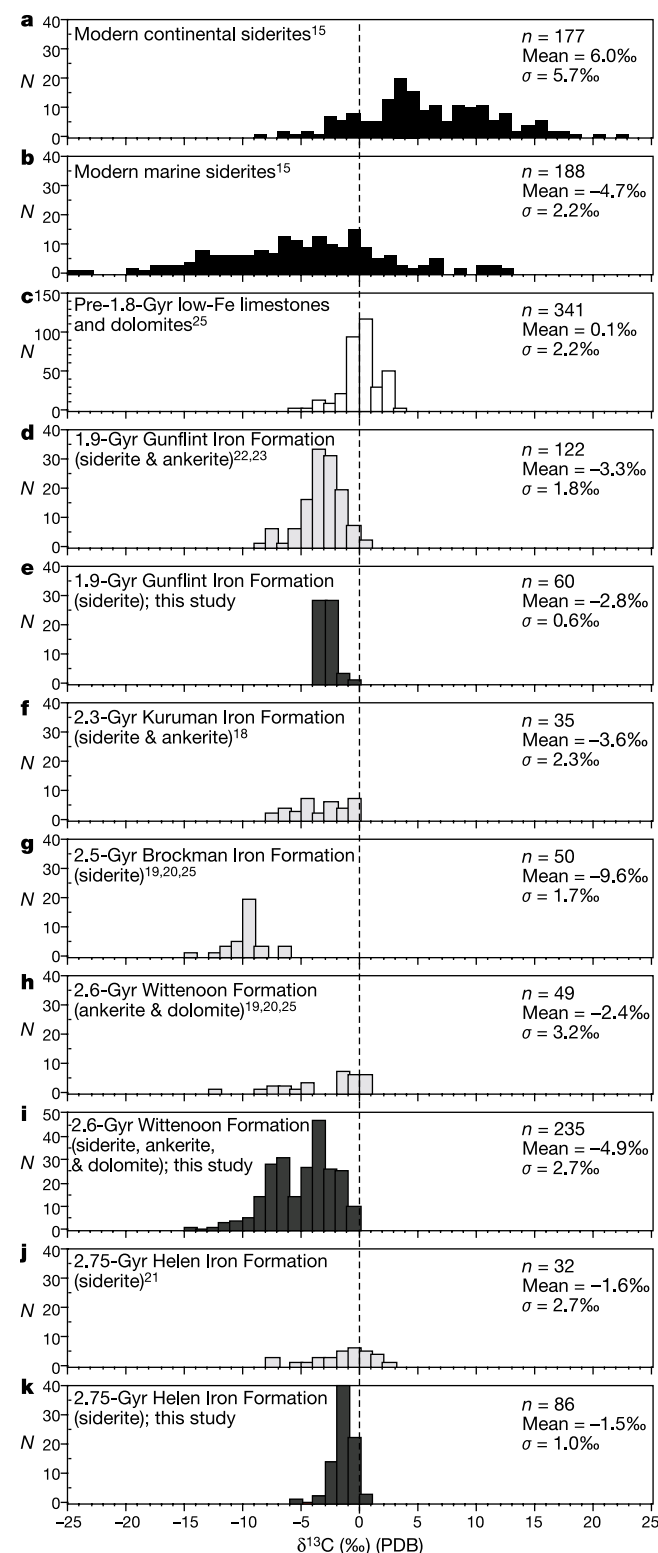


Figure 3 Comparison of the carbon isotopic compositions ($\delta^{13}\text{C}$ values) of carbonates in various rocks. **a, b**, Siderites in modern sediments; **c**, pre-1.8-Gyr Fe-poor (<1 wt% Fe) limestones and dolomites; **d–k**, pre-1.8-Gyr massive siderite-rich sedimentary rocks. We note that the $\delta^{13}\text{C}$ values of pre-1.8-Gyr massive siderite beds are distinctly different from those of modern siderites that formed in terrestrial environments where a significant fraction of the ΣCO_3^{2-} was produced by methanogens. The $\delta^{13}\text{C}$ data on pre-1.8-Gyr massive siderite beds can best be explained if ~ 70 to $\sim 95\%$ of the ΣCO_3^{2-} used in the siderite formation were derived from normal ocean water and the remaining ΣCO_3^{2-} from the decomposition of marine organic matter by non-methanogenic microbes.

tigators^{18–23}: (1) pre-1.8-Gyr massive siderite-rich beds are mostly composed of Fe-rich carbonates and chert (silica-rich chemical precipitates) with very minor components of detrital minerals (for example, clays, feldspars); (2) siderite is typically the earliest carbonate mineral in these beds, often followed by ankerite, dolomite and calcite; (3) the $\delta^{13}\text{C}$ values of pre-1.8-Gyr Fe-poor carbonates largely fall within a range of $0 \pm 2\text{‰}$ (Fig. 3c), which is also the typical range for normal marine carbonates of younger geologic age^{24,25}; however, those of pre-1.8-Gyr Fe-rich carbonates are generally negative (Fig. 3d–k). From $\delta^{13}\text{C}$ values of around -5‰ for siderites in the Kuruman Iron Formation (Fig. 3f), Beukes *et al.*¹⁸ suggest that hydrothermal fluids provided CO_2 , as well as Fe^{2+} , for the siderites. However, this hydrothermal model does not explain the frequent occurrences of massive siderite beds with more negative $\delta^{13}\text{C}$ values (for example, $< -10\text{‰}$ in the Brockman Iron Formation) or near 0‰ values (for example, the Helen Iron Formation).

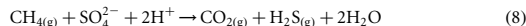
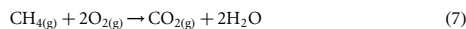
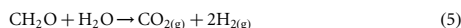
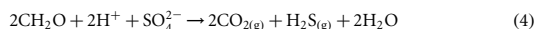
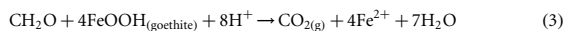
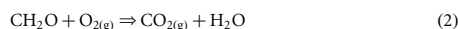
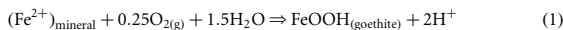
We also agree with the following interpretations made by previous investigators^{19–23} on the above characteristics of pre-1.8-Gyr siderite-rich sedimentary rocks: (1) most siderite crystals directly precipitated from overlying sea water in restricted basins, rather than from pore waters during sediment diagenesis; and (2) the ΣCO_3^{2-} in basin water was a mixture of normal sea water ΣCO_3^{2-} ($\delta^{13}\text{C} = 0 \pm 2\text{‰}$) and biogenic ΣCO_3^{2-} ($\delta^{13}\text{C} = -30 \pm 5\text{‰}$) generated from the decomposition of organic matter ($\delta^{13}\text{C}$ typically between -35 and -20‰) by aerobes and non-methanogenic anaerobes (see reactions (3)–(5) in Methods). The $\delta^{13}\text{C}$ values of pre-1.8-Gyr massive siderite beds (Fig. 3d–k) suggest that the proportion of these two types of ΣCO_3^{2-} varied with time (stratigraphy). These variations were probably due to changes in microbial activity during the history of sedimentary basins; yet the water body in each basin had a characteristic range of $\delta^{13}\text{C}$ values at any given time. Simple isotope mass balance calculations suggest that the percentage of ΣCO_3^{2-} from normal sea water ($\delta^{13}\text{C} \approx 0\text{‰}$) in the mixtures varied from $\sim 70\%$ for the Brockman Iron Formation to $\sim 95\%$ for the Helen Iron Formation (Fig. 3d–k). Findings of organic matter with very low $\delta^{13}\text{C}$ values ($\sim -50\text{‰}$) in some ~ 2.7 -Gyr sedimentary rocks in the Hamersley Basin led Hayes²⁶ to suggest a basin environment where methanogens and methane oxidizers (methanotrophs and/or sulphate reducers) were actively producing biogenic $\text{CO}_{2(\text{aq})}$ with $\delta^{13}\text{C} \ll -30\text{‰}$ (see reactions (6) and (7) in Methods). If such CO_2 was involved in the formation of siderites in the Brockman Iron Formation, the percentage of normal sea water ΣCO_3^{2-} would have been $\geq 70\%$.

The above discussion leads us to conclude that the ΣCO_3^{2-} source for pre-1.8-Gyr siderites was principally contemporaneous sea water. Widespread, thick siderite-rich beds before ~ 1.8 Gyr ago occur because the ΣCO_3^{2-} concentration of normal oceans was $\geq 10^{-2.5} \text{ m}$ and the pH was between ~ 5.5 and ~ 7.5 ; these were consequences of an atmospheric $p_{\text{CO}_2} > 10^{-1.4 \pm 0.2} \text{ atm}$ and a rainwater pH < 4.4 (Fig. 2b). This minimum p_{CO_2} value is more than an order of magnitude higher than the p_{CO_2} values of Pavlov *et al.*^{4,5,27} (for example, $10^{-2.6} \text{ atm}$ at 2.75 Gyr ago), further implying that the atmospheric p_{CH_4} was much lower than the values they estimated (Fig. 1). Our minimum p_{CO_2} value is compatible with the p_{CO_2} range estimated by Kasting⁶ (for example, $10^{-1.2}$ to $10^{+0.2} \text{ atm}$ at 2.75 Gyr ago; $10^{-1.8}$ to $10^{-0.7} \text{ atm}$ at 2.0 Gyr ago; Fig. 1), based on the assumption that CO_2 was the only major greenhouse gas. However, in order to increase the accuracy in estimating atmospheric p_{CO_2} values during the Archaean and Proterozoic eras, we must obtain accurate information on the global climate change, as well as the chemical and isotopic characteristics of Fe-rich carbonates of various age. Our study highlights the need to investigate the impacts of higher atmospheric p_{CO_2} , higher oceanic ΣCO_3^{2-} , lower rainwater pH and lower oceanic pH on global geochemical cycles and on the biota in the oceans and on land. □

Methods

Controls on p_{O_2} and p_{CO_2} values of subsurface waters

The following reactions involving Fe-bearing minerals and organic matter (CH_2O) play important roles in controlling the p_{O_2} and p_{CO_2} of subsurface waters:



Reactions (1) and (2) are typically carried out by aerobic organisms while reactions (3)–(6) are carried out by anaerobes: reaction (3) by Fe^{3+} reducers, reaction (4) by sulphate reducers, reaction (5) by fermenters and reaction (6) by methanogens. The oxidation of CH_4 to CO_2 may be carried out by methanotrophs in reaction (7) and/or sulphate reducers in reaction (8).

According to Henry's law constant for O_2 (ref. 28) and Avogadro's number, the concentration of dissolved O_2 ($mO_{2(aq)}$) in water is $10^{-2.90}$ moles per kg H_2O ($= 10^{20.88}$ molecules per kg H_2O) when the water is in equilibrium with an atmosphere at $p_{O_2(g)} = 1$ atm and $T = 25^\circ C$. Free O_2 molecules disappear from a kilogram of water at $mO_{2(aq)} = 10^{-23.78}$ (that is $p_{O_2(g)} \approx 10^{-21}$ atm) (Fig. 2a). Below this value, the $mO_{2(aq)}$ (and $p_{O_2(g)}$) are 'virtual values' that are 'calculated' from the number of $H_{2(aq)}$ molecules by assuming equilibrium in the reaction: $H_{2(aq)} + 1/2O_{2(aq)} = H_2O$. Therefore, reactions (1) and/or (2) alone may reduce the p_{O_2} of subsurface waters to $\sim 10^{-21}$ atm, but H_2 generation by reaction (5) is required to decrease the p_{O_2} to $\leq 10^{-60}$ atm where Fe^{2+} -rich silicates (for example, minnesotaite) and/or siderite may form. Microbial decomposition of organic matter (reactions (2)–(6)) may also create the other conditions necessary for siderite formation in modern sediments: increases in p_{CO_2} from $\sim 10^{-3.5}$ to $>10^{-1.4}$ atm and in ΣCO_3^{2-} from $\sim 10^{-3.2}$ to $>10^{-2.5}$ m. The increased p_{CO_2} may then decrease pH from 8.1 to <7.5 , mainly through the reaction $CO_{2(g)} + H_2O = H^+ + HCO_3^-$. Organic acids, generated from organic matter decomposition, also contribute to lowering the pH and also play an important role in leaching iron from rocks²⁸ to increase the concentration of ΣFe^{2+} in pore waters from $\ll 10^{-6}$ to $>10^{-6}$ m (reaction 3).

Mineralogical and geochemical investigations of siderite-rich rocks

To confirm some of the major suggestions made by previous investigators^{18–23} on the origins of pre-1.8-Gyr massive siderite beds, we have carried out our own investigations, focusing on microscale variability in mineralogical, chemical and isotopic characteristics of three major siderite-rich sedimentary formations. We selected four specimens of the 2.7-Gyr Helen Iron Formation in the Michipicoten district, Ontario, Canada, 45 specimens of the 2.6-Gyr Wittenoon Formation in the Hamersley district, Western Australia and three specimens of 1.9-Gyr Gunflint Formation in the Thunder Bay district, Ontario, Canada.

We have examined the textural relationships of various carbonates in thin sections from these samples using petrographical and electron microscopes. From each of the specimens, 1 to 30 powdered samples were microdrilled, and a total of 381 powdered samples were analysed for: (1) the abundance ratios of minerals (siderite, ankerite, dolomite, calcite, quartz and so on) using an X-ray diffractometer; (2) the abundance ratios of Fe, Mn, Mg and Ca using an atomic absorption spectrometer; and (3) the $\delta^{13}C$ and $\delta^{18}O$ values using a Finnigan 242 mass spectrometer. The contents and $\delta^{13}C$ values of kerogen in the powdered samples were also determined using a Carlo Erba elemental analyser and the mass spectrometer, respectively. The chemical and isotopic data are presented in the Supplementary Information.

Received 27 December 2003; accepted 16 April 2004; doi:10.1038/nature02573.

- Rye, R., Kuo, P. & Holland, H. D. Atmospheric carbon dioxide concentrations before 2.2 billion years ago. *Nature* **378**, 603–605 (1995).
- Rye, R. & Holland, H. D. Paleosols and the evolution of atmospheric oxygen: A critical review. *Am. J. Sci.* **298**, 621–672 (1998).
- Catling, D., Zahnle, K. & McKay, C. Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* **293**, 839–843 (2001).
- Pavlov, A. A., Brown, L. L. & Kasting, J. F. UV shielding of NH_3 and O_2 by organic hazes in the Archean atmosphere. *J. Geophys. Res.* **106**, 23267–23287 (2001).
- Pavlov, A. A., Kasting, J. F., Eigenbrode, J. L. & Freeman, K. H. Organic haze in Earth's early atmosphere: Source of low- ^{13}C Late Archean kerogens? *Geology* **29**, 1003–1006 (2001).
- Kasting, J. Theoretical constraints on oxygen and carbon dioxide concentrations in the Precambrian atmosphere. *Precamb. Res.* **34**, 205–229 (1987).
- Ohmoto, H. Evidence in pre-2.2 Ga paleosols for the early evolution of atmospheric oxygen and terrestrial biota. *Geology* **24**, 1135–1138 (1996).
- Langmuir, D. *Aqueous Environmental Geochemistry* 600 (Prentice-Hall, Upper Saddle River, 1997).
- Panahi, A., Young, G. M. & Rainbird, R. H. Behavior of trace elements (including REE) during pedogenesis and diagenetic alteration of an Archean granite near Ville Marie, Québec, Canada. *Geochim. Cosmochim. Acta* **64**, 2199–2220 (2000).
- Marmo, J. S. in *Early Organic Evolution* (ed. Schidlowski, M.) 41–66 (Springer, Berlin, 1992).
- Beukes, N. J., Dorland, H., Gutzmer, J., Nedachi, M. & Ohmoto, H. Tropical laterites, life on land, and

the history of atmospheric oxygen in the paleoproterozoic. *Geology* **30**, 491–494 (2002).

- Watanabe, Y., Stewart, B. W. & Ohmoto, H. Organic- and carbonate-rich soil formation ~2.6 billion years ago at Schagen, East Transvaal district, South Africa. *Geochim. Cosmochim. Acta* **68**, 2129–2151 (2004).
- Knauth, L. P. & Lowe, D. R. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. *Geol. Soc. Am. Bull.* **115**, 566–580 (2003).
- Moore, S. E., Ferrell, R. E. & Aharon, P. Diagenetic siderite and other ferroan carbonates in a modern subsiding marsh sequence. *J. Sedim. Petrol.* **62**, 357–366 (1992).
- Mozley, P. & Wersin, P. Isotopic composition of siderite as an indicator of depositional environment. *Geology* **20**, 817–820 (1992).
- Kimberley, M. M. Exhalative origins of iron formations. *Ore Geol. Rev.* **5**, 13–145 (1989).
- Ohmoto, H. When did the atmosphere become oxic? *Geochim. News* **93**, 12–13, 26–27 (1997).
- Beukes, N. J., Klein, C., Kaufman, A. J. & Hayes, J. M. Carbonate petrography, kerogen distribution, and carbon and oxygen isotope variations in an early Proterozoic transition from limestone to iron-formation deposition, Transvaal Supergroup, South Africa. *Econ. Geol.* **85**, 663–690 (1990).
- Becker, R. H. & Clayton, R. N. Carbon isotope evidence for the origin of a banded iron-formation in western Australia. *Geochim. Cosmochim. Acta* **36**, 577–595 (1972).
- Baur, M. E., Hayes, J. M., Studley, S. A. & Walter, M. R. Millimeter-scale variations of stable isotope abundances in carbonates from banded iron-formations in the Hamersley Group of Western Australia. *Econ. Geol.* **80**, 270–282 (1985).
- Goodwin, A. M., Thode, H. G., Chou, C.-L. & Karkhans, S. N. Chemostratigraphy and origin of the late Archean siderite-pyrite-rich Helen Iron Formation, Michipicoten belt, Canada. *Can. J. Earth Sci.* **22**, 72–84 (1985).
- Carrigan, W. J. & Cameron, E. M. Petrological and stable isotope studies of carbonate and sulfide minerals from the Gunflint Formation, Ontario: evidence for the origin of early Proterozoic iron-formation. *Precamb. Res.* **52**, 347–380 (1991).
- Winter, B. L. & Knauth, P. Stable isotope geochemistry of cherts and carbonates from the 2.0 Ga Gunflint Iron Formation: implications for the depositional setting, and the effects of diagenesis and metamorphism. *Precamb. Res.* **59**, 283–313 (1992).
- Schidlowski, M. et al. in *Early Organic Evolution* (ed. Schidlowski, M.) 147–175 (Springer, Berlin, 1992).
- Shields, G. & Veizer, J. Precambrian marine carbonate isotope database: Version 1.1. *Geochim. Geophys. Geosyst.* **3**, 1–12 (2002).
- Hayes, J. M. in *Early Life on Earth* (ed. Bengtson, S.) 220–236 (Columbia Univ. Press, New York, 1994).
- Pavlov, A. A., Hurtgen, M. T., Kasting, J. F. & Arthur, M. A. Methane-rich Proterozoic atmosphere? *Geology* **31**, 87–90 (2003).
- Stumm, W. & Morgan, J. J. *Aquatic Chemistry, Chemical Equilibria and Rates in Natural Waters* 3rd edn 1022 (Wiley, New York, 1996).
- Geochemist's Workbench* Version 4 (RockWare Inc., Golden, 2002).
- Ohmoto, H. Redox state of the Archean atmosphere: Evidence from detrital heavy minerals in ca. 3250–2750 Ma sandstones from the Pilbara Craton, Australia. *Geology* **27**, 1151–1152 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to K. Hayashi and T. Kakegawa for assistance in petrographical and geochemical investigations of the siderite samples. We thank M. Arthur, H. Barnes, E. Bazilevskaya, P. Deines, L. Kump, A. Lasaga, K. Spangler and K. Yamaguchi for comments on earlier drafts. This work was supported by grants to H.O. from the NSF (Geochemistry Program) and NASA (Astrobiology and Exobiology programmes).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.O. (ohmoto@geosc.psu.edu).

Low electrical resistivity associated with plunging of the Nazca flat slab beneath Argentina

John R. Booker¹, Alicia Favetto² & M. Cristina Pomposiello²

¹Department of Earth and Space Sciences, Box 351310, University of Washington, Seattle, Washington 98195, USA

²Instituto de Geocronología y Geología Isotópica, Ciudad Universitaria, Pabellón INGEIS, 1428 Buenos Aires, Argentina

Beneath much of the Andes, oceanic lithosphere descends eastward into the mantle at an angle of about 30° (ref. 1). A partially molten region is thought to form in a wedge between this descending slab and the overlying continental lithosphere as volatiles given off by the slab lower the melting temperature of

mantle material². This wedge is the ultimate source for magma erupted at the active volcanoes that characterize the Andean margin. But between 28° and 33° S the subducted Nazca plate appears to be anomalously buoyant^{3,4}, as it levels out at about 100 km depth and extends nearly horizontally under the continent^{1,5,6}. Above this 'flat slab', volcanic activity in the main Andean Cordillera terminated about 9 million years ago as the flattening slab presumably squeezed out the mantle wedge^{5,6}. But it is unknown where slab volatiles go once this happens, and why the flat slab finally rolls over to descend steeply into the mantle 600 km further eastward. Here we present results from a magnetotelluric profile in central Argentina, from which we infer enhanced electrical conductivity along the eastern side of the plunging slab, indicative of the presence of partial melt. This conductivity structure may imply that partial melting occurs to at least 250 km and perhaps to more than 400 km depth, or that melt is supplied from the 410 km discontinuity, consistent with the transition-zone 'water-filter' model of Bercovici and Karato⁷.

Contours of the depth of seismic activity associated with the Nazca slab under central Chile and Argentina are shown in Fig. 1. The absence of Andean volcanic activity where the subducted slab is nearly flat is evident, as is the spatial coincidence of the Sierras Pampeanas with the extent of the flat slab. The Sierras Pampeanas are a series of roughly north–south ridges separated by wide desert valleys east of the Andes. They consist of crystalline basement blocks that have been uplifted along faults that dip steeply into the deep crust^{6,8}. This 'thick-skinned' deformation is still active and coincides in time with the presence of the flat slab. It is reactivation of weaknesses formed during the series of continental collisions and break-ups that over the last 600 million years (Myr) added the western half of South America to the much older Rio de la Plata craton^{6,9}. In the west, the thin skin of the Andean Pre-Cordillera is thrust eastward over the Sierras Pampeanas basement. The easternmost ridges of the Sierras Pampeanas, the Sierras de Córdoba, coincide with the edge of the craton and the eastern extent of the Nazca flat slab⁶.

The magnetotelluric (MT) method uses surface measurements of natural fluctuations of electric and magnetic fields generated in the

atmosphere and magnetosphere to infer structure of electrical conductivity (σ) or resistivity ($\rho = 1/\sigma$) of Earth's interior¹⁰. Silicate minerals in the crust (0 to about 40 km) and upper mantle (to 410 km) are generally electrical insulators¹¹. Low ρ is due to interconnected grain boundary phases such as water, magma and in some instances, metals, sulphides or graphite^{12,13}. In the transition zone between 410 and 660 km, hydrogen solubility is higher¹⁴ and enhanced conductivity has been observed in minerals thought to exist in this depth range¹⁵. Deeper than 660 km, the dominant minerals are intrinsically more conductive^{15,16}. Crustal models are usually presented as resistivity (Ω m), whereas mantle models are often given as conductivity ($S\ m^{-1}$).

The same physics that increases electrical transport properties, such as enhanced diffusion and fluid-filled cracks, usually makes rocks more deformable¹⁷. Crystalline basement such as the Sierras Pampeanas and Rio de la Plata craton⁶ would normally be electrically resistive and strong. Furthermore, in most areas, the Sierras Pampeanas and the plains east of the Sierras de Córdoba are covered with only a thin veneer of conductive sediment and sedimentary rocks¹⁸. The situation is favourable for deep penetration of electromagnetic energy and makes MT an appropriate tool for investigating the physical state of the upper mantle.

This Letter reports interpretation of an MT profile across the eastern edge of the Sierras Pampeanas above the widest extent of the Nazca flat slab. Figures 1 and 2 show the location of the 18 sites with data from 2.5 to 7,000 s period. The data are frequency (f) domain transfer functions between the horizontal electric vector (E) and the horizontal magnetic vector (H): $E = ZH$, where Z is known as the MT impedance tensor and between the vertical magnetic component and the horizontal magnetic vector: $H_z = TH$, where T is called the induction vector. Each of the four complex elements of Z has a phase (ϕ) and a magnitude $|Z|$, which is usually rescaled as an apparent resistivity, $\rho_a = |Z|^2/5f$.

It is difficult to interpret a single MT profile unless one can justify assuming two-dimensional (2D) structure. Geology and tectonics supports a dominant north–south strike (Fig. 1). MT and induction-vector data also contain information about dimensionality and strike¹⁹. In the Supplementary Information we conclude that if error estimates in the complex elements of Z from 5 to 5,000 s are increased to the equivalent of $\pm 10\%$ in ρ_a ($\pm 2.9^\circ$ in ϕ) and from 10 to 3,600 s are increased to ± 0.03 in the real and imaginary parts of T , these data are compatible with 2D and a north–south regional strike. These inflated errors exceed the statistical error estimates at all periods. We normalize misfit of the predicted responses to the measured data by the inflated error estimates so that a normalized r.m.s. misfit (n.r.m.s.) of 1.0 corresponds to fitting the data, on average, to the level just consistent with the 2D assumption.

When the structure is 2D, the data separate into two independent modes¹⁰ with electric current flowing parallel and perpendicular to strike. The TM (E -perpendicular) mode corresponds to electric currents flowing east–west along the profile and across the structural strike, and the TE (E -parallel) mode represents currents flowing north–south across the profile and along the strike. Induction-vector data are an important aspect of the TE mode because they can sense lateral structural variations beyond the ends of the profile. The data are inverted using a 2D algorithm²⁰ that seeks minimum structure or 'roughness' of $\log(\rho)$ (that is $-\log(\sigma)$) for given data misfit. We are thus seeking structure required rather than allowed by the data. TM ρ_a data were down-weighted relative to ϕ because spatial under-sampling of the response disturbs ρ_a more than ϕ . TE ρ_a data were not used at all because they are most susceptible to distortion by off-profile structure. The penalty on both the vertical and horizontal roughness increases with depth and with distance outside the profile. This prejudices the inversion to find shallow, local structure if possible. Initial inversions consistently required large lateral gradients deeper than 200 km. We therefore increased the penalty on deep lateral variation by

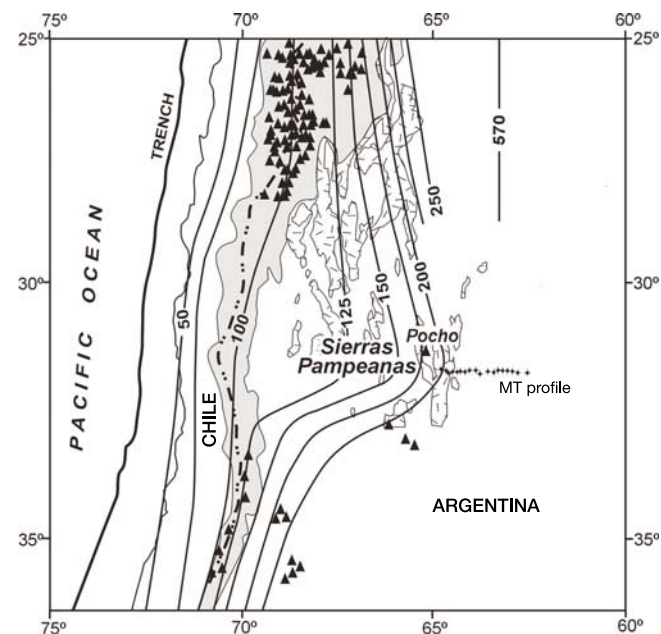


Figure 1 Location map showing main physiographic features, contours of the depth in kilometres to the Wadati–Benioff earthquake zone¹ and the MT sites. The grey area is the Andean Cordillera above 3,000 m; the stippled areas are the Sierra Pampeanas block uplifts; filled triangles are volcanoes active in the last 8 Myr.

constraining ρ deeper than 660 km to a value reflecting the expected global decrease^{15,16}. The actual ρ for this constraint has little other influence as long as the contrast with the upper mantle is large. We used $3 \Omega \text{ m}$ (0.3 S m^{-1}).

Figure 2a is the trade-off curve of n.r.m.s. versus model roughness. It has three distinct segments. Along the blue segment n.r.m.s. drops rapidly while roughness increases slowly. Along the red segment, n.r.m.s. decreases slowly while roughness increases rapidly and models develop small-scale structure that is increasingly unlikely as the misfit drops below that consistent with 2D interpretation. Figure 2b,c presents the models at the transitions from the blue and red segments to the intervening green segment. Model b (shown in Fig. 2d) is a pessimistic view of the information in our data, whereas models fitting more tightly than model c (shown in Fig. 2c) may be fitting 3D aspects of the data with 2D structure and are thus too optimistic. Figure 2d expands the shallow part of model c. Models along the entire trade-off curve are shown in the Supplementary Information. Also shown in the Supplementary Information are comparisons of measured data and computed responses for model c.

On the green segment, model focus sharpens as n.r.m.s. decreases, but gross features persist. Thin sedimentary cover overlies highly resistive upper crust except in the Sierras de Córdoba where crystalline basement rocks are exposed at the surface. The lower crust below 20 km is conductive directly under the highest elevation of the Sierras de Córdoba, but is resistive to the east. The most surprising feature is a conductive zone rising from deeper mantle to about 100 km under our profile. East of this vertical conductor, high resistivity extends to more than 200 km depth. Resistive mantle

west of the vertical conductor extends deeper, but is not as resistive. Because the mantle deeper than 660 km is constrained to be conductive by mineralogical considerations, the minimum structure inversion pushes the conductive mantle up as far as data misfit allows. This is why the bottom of the resistive mantle gets deeper as n.r.m.s. decreases.

Earthquake hypocentres from the Advanced National Seismic Systems (ANSS) composite catalogue with an estimated magnitude of 5.0 or more from 1973–2002 are plotted on the models of Figs 2b, c, e. Shallower than 200 km, these are the events with well-constrained depths between 30 and 32° S. The solid black line is the depth of the Nazca slab inferred from these events. Below 200 km the slab is shown dashed because there are no deeper earthquakes at this latitude. Earthquakes deeper than 500 km terminate abruptly at 29° S (Fig. 1). The events deeper than 300 km and south of 28° S in the same time interval are also plotted on the model and we have extrapolated the position of the deep slab south to our profile under the assumption that the well-defined north–south trend of the deep events continues. The remarkable correspondence between the vertical mantle conductor and the predicted position of the deep Nazca slab strongly supports both this extrapolation, and that the origin of the conductor is related to the presence of the slab.

High resistivity extending to at least 200 km east of the vertical conductor can be interpreted as the root of the Rio de la Plata craton. Its great thickness suggests that the descent of the buoyant Nazca flat slab was triggered by collision with this electrically and mechanically resistive root. It probably continues to block development of a normal asthenospheric wedge in mantle flow entrained by the descending slab.

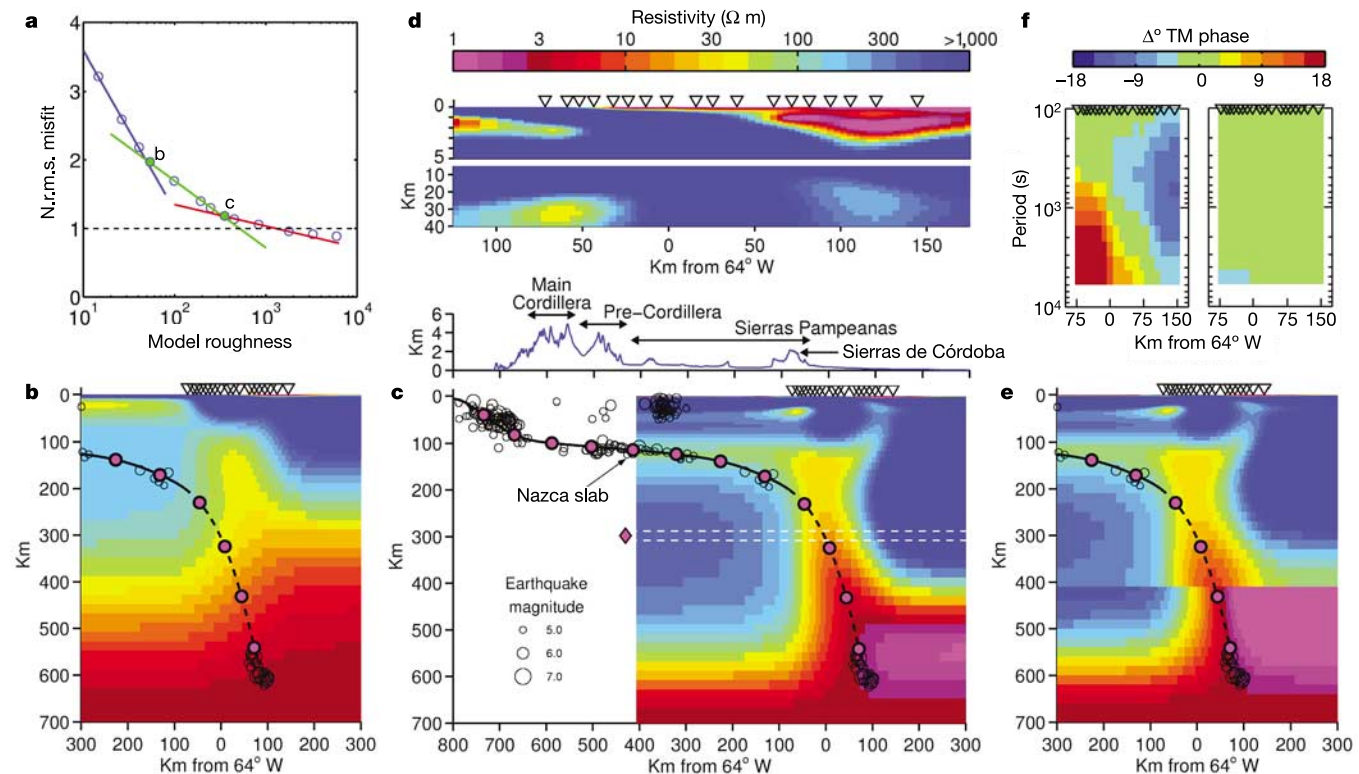


Figure 2 Inversions of MT data for electrical structure beneath central Argentina. Resistivity deeper than 700 km is constrained to $3 \Omega \text{ m}$. The position of the Nazca slab is estimated from plotted earthquake hypocentres and is shown with 1-Myr isochrons since subduction plotted as filled circles. Inverted triangles show MT site locations. **a**, Trade-off of data misfit with model roughness. Misfit is normalized by the minimum permitted for 2D interpretation. **b**, Model at trade-off point b in Fig. 2a. **c**, Model at trade-off point c in Fig. 2a. **d**, Expanded crustal portion of model c. Upper panel is 0 to 5 km with a 6:1 vertical exaggeration. Lower panel is 5 to 40 km with no vertical exaggeration. **e**, Model with same

n.r.m.s. as model c but with no penalty on structure across a 410 km discontinuity. **f**, TM phase differences as a function of period and site position for two hypothesis tests. The central colour of the bar is $\pm 3^\circ$, the error estimate used (see text). Left: effect of adding the resistive layer indicated by horizontal dashed lines and a diamond to model c; maximum difference exceeds six times the error. Right: difference between model e and the model with resistivity constrained to $3 \Omega \text{ m}$ below a 410-km discontinuity; at most periods the difference does not exceed the error estimate; the maximum difference is twice the error estimate.

The effective electromagnetic skin depth in the east of model c is 400 km at 200 s period and 500 km at 5,000 s. To the west, the effective skin depth is 400 km at 1,700 s and 590 km at 5,000 s. Thus, energy within our measurement band easily penetrates to the top of the horizontal deep conductor, but not far into it. One does not usually expect to resolve lateral structure deeper than the length of an MT profile. It is therefore reasonable to ask whether the vertical connection between the conductor near 100 km under our profile and the conductive deeper mantle is required. Resolution tests common in seismology such as inversion of data generated by a checkerboard model are not helpful because diffusive electromagnetic responses depend on the entire electric current path. However, a simple test demonstrates that our data are sensitive to deep structural details due to favourable geometry. A thin resistive layer inserted at 300 km across model c cuts the electrical path from the deep mantle in the east to the shallower conductors in the west. This increases n.r.m.s. from 1.18 to 2.0. TM ϕ changes by more than ± 3 times the inflated error estimate over much of the profile (see Fig. 2e). The TM data (which are least affected by 3D complications²¹) clearly require that this path be continuous. Thus the vertical current path is a very robust feature. At depth it coincides more closely with the slab as n.r.m.s. decreases.

Above 200 km, the vertical conductor rises above the Nazca slab until it seems to meet an impermeable boundary at 100 km. Model c completes the required electrical circuit to the west by penetration of low resistivity between the flat-slab and this cap. Model b does not resolve this structure and completes the circuit by smearing the lower crustal conductor under the Sierras de Córdoba vertically and to the west. The subtle structure in model c rising from the eastern end of the vertical mantle conductor to about 10 km under the eastern sedimentary basin can be removed without affecting data misfit. It is therefore not reliable evidence for upward penetration of the conductive medium. We conclude that the Sierras Pampeanas lithosphere is a 100-km-thick 'bridge' kept under compression between the craton and the Andes. This compression prevents significant upward penetration of slab fluids, partial melt and heat advection. The present geometry has not existed long enough for the thermal signature of processes at 100 km to diffuse to the surface by conduction. Thus the Sierras Pampeanas lithosphere remains largely cold, electrically resistive and strong.

Model b suggests the deep conductivity increase coincides with the 410-km seismic discontinuity west of the vertical conductor, whereas model c suggests that this gradient coincides with this transition on the east side. As laboratory data also suggest a sharp increase in conductivity is possible at the 410-km phase transition¹⁵, we can ask whether a model with a jump in conductivity at 410 km across the entire model is compatible with our data. We repeated the entire suite of inversions with the model constrained to 3 Ω m below 410 km and a roughness penalty that ignores any gradients across this discontinuity. This new suite of constrained inversions has a trade-off curve almost identical to Fig. 2a. In particular, it has the two slope breaks labelled b and c in Fig. 2a. We refer to the models at the two slope breaks for this new suite as 'corresponding' to models b and c.

The model corresponding to b differs from model b in having the steep vertical gradient in the west collapse to the imposed 410-km discontinuity. The region of steep vertical gradient in the east is near 250 km in both models. Thus at an n.r.m.s. of 2, the discontinuity at 410 km and the constrained resistivity deeper than 410 km do not affect the depth to the high conductivity east or west of the slab.

The model corresponding to c differs little from model c above 410 km. However, the n.r.m.s. of the constrained model is slightly higher than model c. We therefore relaxed the constraint of fixed resistivity below 410 km and continued to iterate the inversion starting from the model corresponding to c. The inversion converged to the model shown in Fig. 2e. Its n.r.m.s. is now the same as model c. The transition zone to the east of the descending slab has become more conductive and the transition zone to the west has

become more resistive. There is no longer a significant conductivity jump at the imposed 410 km discontinuity to the west. Figure 2f (right) shows the difference in TM ϕ response between model e and the model corresponding to c (that is, with 3 Ω m below 410 km). The phase change exceeds the inflated error estimate only at the longest periods and its maximum is only about twice the inflated error estimate. So we cannot rule out a flat discontinuity in ρ at 410 km. However, this phase difference is consistent across the western third of our profile and the roughly 150-km step is a feature of models with much higher n.r.m.s. than models c and e. We conclude that the existence of the step is likely, but its depth is poorly constrained.

Resistivity $\leq 30 \Omega$ m deeper than 200 km rules out graphite²² and the solid-state conductivity of mantle minerals shallower than 410 km^{12,15} as significant contributors to the vertical mantle conductor. Interconnected aqueous fluids or melt are alternatives. The observed resistivities require less than 5% melt if it is well connected^{13,23,24}. Hydrous fluids can be more conductive and require less interconnected fluid. The convergence of the Nazca plate with respect to stable South America at our profile has declined from about 140 to 65 km per Myr over the last 26 Myr²⁵. The filled circular points along the Nazca slab in Fig. 2 are 1-Myr isochrons since subduction based on this estimate. Unless convergence rates were much slower than believed, the slab presently at 400 km depth was west of the Andes when volcanism ceased 8.6 Myr ago and west of the most eastern volcanic centre, Pocho (Fig. 1) when volcanism ceased 4.5 Myr ago^{5,6}. Consequently, the entire plate from 400 km depth up to the surface has never been below active volcanism and may retain significant volatiles. Cratonic mantle is too hot below about 200 km to host hydrous fluids coming from the slab without melting^{3,26}. At these depths, hydrous fluids are completely miscible with melt²⁷ and cannot exist independently. Thus the only situation in which the vertical conductor can be due entirely to hydrous fluids is if they are confined to a cool slab that is internally permeable. A sealed yet permeable slab seems unlikely. Partial melting of the mantle as fluids leak from the slab is preferred and a combination of partial melt and fluids in the adjacent slab seems the most reasonable explanation for enhanced conductivity below 250 km where the conductor coincides most closely with the slab. However, the required vertical continuity of the conductor implies that breakdown of hydrous minerals in the slab to flux this melting must continue downwards until the vertical conductor connects with the more widespread deep mantle conductivity, that is, to at least 250 km and perhaps into the transition zone deeper than 410 km.

An alternative view²⁸ of the fate of flat-slab volatiles maintains that significant amounts may be lost to the overlying lithosphere before the flat slab starts to plunge. Melt is not produced because the lithospheric temperature is too low. However, massive melting of wet lithosphere could explain the commonly observed flare-up in volcanism associated with reestablishment of the asthenospheric wedge when the flat slab returns to a more normal dip. The wet lower lithosphere also provides an ideal candidate for the required shallow electric-current connection to the west, but if the fluids are from the flat slab, this implies a relatively dry plunging slab less likely to flux significant partial melting of adjacent mantle below 200 km. Is there another explanation for the vertical conductor that avoids the uncomfortable necessity of transporting more than trace amounts of slab fluids to great depth?

Bercovici and Karato⁷ (B-K) have put forward a provocative hypothesis to explain geochemical differences between basaltic lavas erupted at mid-ocean ridges (MORB) and those erupted at oceanic islands such as Hawaii (OIB). The traditional view is that MORB is derived by partial melting of mantle above 410 km that is isolated from a deeper source for OIB. A difficulty with this view is the considerable evidence that return flow from the surface penetrates into the deep mantle. Conservation of mass requires upward flow through the transition zone in excess of that associated with OIB

volcanism. B–K argue that this vertical transport is widespread and of such low amplitude that it has escaped notice, but it makes isolation of geochemical reservoirs above and below 410 km difficult to maintain. They further argue that hydrogen dissolved in phases below the 410 km discontinuity is released as water when minerals rise through the 410 km transition. They suggest that this results in partial melting at the discontinuity and that, unlike melt produced in shallow asthenospheric wedges, this melt is dense enough that it does not rise into the upper mantle. The melt, however, strips out the elements necessary to convert the OIB source to MORB source.

The mantle for at least several hundred kilometres west of the plunging Nazca slab descends with the slab. If the B–K hypothesis is correct, it would have been stripped of water when it rose through 410 km under the East Pacific Rise. This mantle should be more resistive as it descends below 410 km than it was when it left the transition zone. East of the slab is the distal limb of the Mid-Atlantic Ridge. Even if there is little or no upward flux through 410 km close to the plunging slab, the minerals below 410 km should not have been stripped of hydrogen and should be more conductive. This predicts a jump in conductivity at the 410 km discontinuity east of the slab and that the plunging slab should separate conductive from resistive transition zone. Hence the B–K hypothesis predicts transition zone structure in remarkable agreement with models c and e. Furthermore, where the slab comes in contact with hydrogen-rich mantle at 410 km it may take the addition of very little hydrous fluid from the slab to either initiate partial melting or to make existing partial melt at the discontinuity buoyant. Thus an intriguing possibility is that the source of partial melt streaming up the east side of the plunging Nazca slab is the transition zone, and that the melt, if it reached the surface, might resemble OIB. This may explain OIB-like volcanism in subduction zones without invoking deep mantle plumes or other complicated mechanisms.

Our data cannot definitively place the increase of conductivity east of the plunging slab at 410 km or exclude a 410 km increase west of the slab. Accurate long-period data along extensions of our profile are needed. More generally, the B–K hypothesis predicts that areas of broad mantle upwelling should have high conductivity in the transition zone, whereas areas of downwelling should be more resistive. Carefully sited long period MT experiments can test this prediction. □

Received 27 February; accepted 8 April 2004; doi:10.1038/nature02565.

- Cahill, T. & Isacks, B. Seismicity and shape of the subducted Nazca Plate. *J. Geophys. Res.* **97**, 17503–17529 (1992).
- Ulmer, P. Partial melting in the mantle wedge – the role of H₂O in the genesis of mantle-derived “arc-related” magmas. *Phys. Earth Planet. Inter.* **127**, 215–232 (2000).
- Cloos, M. Lithospheric buoyancy and collisional orogenesis: subduction of oceanic plateaus, continental margins, island arcs, spreading ridges and seamounts. *Geol. Soc. Am. Bull.* **105**, 714–737 (1993).
- McGeary, S., Nur, A. & Ben-Avraham, Z. Spatial gaps in arc volcanism: the effect of collision or subduction of oceanic plateaus. *Tectonophysics* **119**, 195–221 (1985).
- Kay, S. M. & Mpodozis, C. Magmatism as a probe of Neogene shallowing of the Nazca Plate beneath the modern Chilean flat-slab. *J. S. Am. Earth Sci.* **15**, 39–57 (2002).
- Ramos, V. A., Cristallini, E. O. & Pérez, D. J. The Pampean flat-slab of the central Andes. *J. S. Am. Earth Sci.* **15**, 59–78 (2002).
- Bercowski, D. & Karato, S. Whole mantle convection and the transition-zone water filter. *Nature* **425**, 39–44 (2003).
- Jordan, T. & Allmendinger, R. Sierras Pampeanas of Argentina: a modern analogue of Rocky Mountain foreland deformation. *Am. J. Sci.* **286**, 737–764 (1986).
- Rogers, J. W. A history of continents in the past three billion years. *J. Geol.* **104**, 91–107 (1996).
- Vozoff, K. in *Electromagnetic Methods in Applied Geophysics* Vol. 2 (ed. Nabighian, M. N.) 641–711 (Society of Exploration Geophysicists, Tulsa, 1991).
- Jones, A. G. in *Continental Lower Crust* (eds Fountain, D. M., Arculus, R. & Key, R. W.) 81–143 (Elsevier, Amsterdam, 1992).
- Jones, A. G. Imaging the continental upper mantle using electromagnetic methods. *Lithos* **48**, 57–80 (1999).
- Schilling, F. R., Partzsch, G. M., Brasse, H. & Schwarz, G. Partial melting below the magmatic arc in the central Andes deduced from geoelectromagnetic field experiments and laboratory data. *Phys. Earth Planet. Inter.* **103**, 17–31 (1997).
- Williams, Q. & Hemley, R. J. Hydrogen in the deep Earth. *Annu. Rev. Earth Planet. Sci.* **29**, 365–418 (2001).
- Xu, Y., Shankland, T. J. & Poe, B. T. Laboratory-based electrical conductivity in the Earth's mantle. *J. Geophys. Res.* **105**, 27865–27875 (2000).
- Peyronneau, J. & Poirier, J. P. Electrical conductivity of the Earth's lower mantle. *Nature* **342**, 537–539 (1989).

- Carter, N. L. & Tsenn, M. C. Flow properties of continental lithosphere. *Tectonophysics* **136**, 27–63 (1987).
- Chebli, G. A., Mozetic, M. E., Rosello, E. A. & Bühler, M. in *Geología Argentina* (ed. Caminos, R.) 627–644 (Inst. de Geología y Recursos Minerales, Buenos Aires, 1999).
- Chave, A. D. & Smith, J. T. On electric and magnetic galvanic distortion tensor decompositions. *J. Geophys. Res.* **99**, 4669–4682 (1994).
- Rodi, W. & Mackie, R. Non-linear conjugate gradient algorithm for 2-D magnetotelluric inversion. *Geophysics* **66**, 174–178 (2001).
- Wannamaker, P. E. in *Three-dimensional Electromagnetics* (eds Oristaglio, M. & Spies, B.) 511–527 (Society of Exploration Geophysicists, Tulsa, 1999).
- Pearson, D. G. *et al.* The characterization and origin of graphite in cratonic lithospheric mantle: a petrological carbon isotope and Raman spectroscopic study. *Contrib. Mineral. Petrol.* **115**, 449–466 (1994).
- Roberts, J. J. & Tyburczy, J. A. Partial-melt electrical conductivity: Influence of melt composition. *J. Geophys. Res.* **104**, 7055–7065 (1999).
- Park, S. & Ducea, M. Can in situ measurements of mantle electrical conductivity be used to infer properties of partial melts? *J. Geophys. Res.* **108**, 2270 (2003).
- Norabuena, E. O., Dixon, T. H., Stein, S. & Harrison, C. G. A. Decelerating Nazca-South America and Nazca-Pacific plate motions. *Geophys. Res. Lett.* **26**, 3402–3404 (1999).
- Wiley, P. J. Magmas and volatile components. *Am. Mineral.* **64**, 469–500 (1979).
- Bureau, H. & Keppler, H. Complete miscibility between silicate melts and hydrous fluids in the upper mantle: experimental evidence and geochemical implications. *Earth Planet. Sci. Lett.* **165**, 187–196 (1999).
- James, D. E. & Sacks, S. in *Geology of Ore Deposits of the Central Andes* (ed. Skinner, B. J.) 1–25 (Special Pub. 7, Society of Economic Geologists, Littleton, Colorado, 1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research would not have been possible without the help of our field technician, G. Giordino of INGEIS and of B. Narod, whose new generation of MT instruments were used to collect the data and who solved critical instrument problems in the field. We also thank M. Lopez, S. Kay, D. James, S. Constable and S.-I. Karato for their insightful comments.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.B. (booker@ess.washington.edu)

Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom

Adrianna Ianora¹, Antonio Miralto¹, Serge A. Poulet², Ylenia Carotenuto¹, Isabella Buttino¹, Giovanna Romano¹, Raffaella Casotti¹, Georg Pohnert³, Thomas Wichard³, Luca Colucci-D'Amato⁴, Giuseppe Terrazzano⁵ & Victor Smetacek⁶

¹Ecophysiology Laboratory, Stazione Zoologica “A. Dohrn”, Villa Comunale, 80121 Naples, Italy

²Station Biologique de Roscoff, CNRS, Place Georges Teissier, 29682 Roscoff, France

³Max Planck Institute for Chemical Ecology, Hans Knöll Strasse 8, D-07745 Jena, Germany

⁴Institute of Endocrinology and Experimental Oncology, CNR, Via Pansini 5, 80131 Naples, Italy

⁵Department of Cellular and Molecular Biology and Pathology, University of Naples, Via Pansini 5, 80131 Naples, Italy

⁶Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

The growth cycle in nutrient-rich, aquatic environments starts with a diatom bloom that ends in mass sinking of ungrazed cells and phytodetritus¹. The low grazing pressure on these blooms has been attributed to the inability of overwintering copepod populations to track them temporally². We tested an alternative explanation: that dominant diatom species impair the reproductive success of their grazers. We compared larval development of a common overwintering copepod fed on a ubiquitous, early-blooming diatom species with its development when fed on a

typical post-bloom dinoflagellate. Development was arrested in all larvae in which both mothers and their larvae were fed the diatom diet. Mortality remained high even if larvae were switched to the dinoflagellate diet. Aldehydes, cleaved from a fatty acid precursor by enzymes activated within seconds after crushing of the cell³, elicit the teratogenic effect⁴. This insidious mechanism, which does not deter the herbivore from feeding but impairs its recruitment, will restrain the cohort size of the next generation of early-rising overwinterers. Such a trans-generational plant–herbivore interaction could explain the recurrently inefficient use of a predictable, potentially valuable food resource—the spring diatom bloom—by marine zooplankton.

The dense phytoplankton blooms that characterize productive regions and seasons in the sea are dominated, from high to low latitudes and from coast line to open ocean, by comparatively few, often cosmopolitan species of diatom⁵. A significant fraction of their biomass sinks out ungrazed, because the populations of their main grazers—calanoid copepods—generally peak well after the bloom, under much lower food concentrations than during the bloom^{6,7}. This paradox has been attributed to the slow response times of copepod life cycles, which involve 11 larval stages and take weeks to months (depending on temperature) to complete⁸. However, the faster-growing protistan grazers of diatoms are also unable to keep track of the bloom. This contrasts strikingly with the ubiquitous microbial food-webs, in which heavy grazing pressure restricts population build-up of all components⁹. Indeed, the bloom-forming diatoms differ from other unicellular plankton in their ability to dominate total pelagic biomass. Clearly, they are less grazed than other species.

Copepod life-cycle strategies differ widely, from year-round reproductive activity to strong seasonality interspersed with dormant stages of varying length⁸. Yet none of the species reaches its annual peak abundance during diatom blooms, suggesting that diatoms are not an ideal food source. A negative effect of diets of various diatom species on copepod egg hatching success (up to 100%) has been demonstrated experimentally^{10,11}, although its ecological implications remain controversial¹². To test whether maternal and post-embryonic diatom diets also affected larval development and survival, we carried out experiments in which pure cultures of the cosmopolitan, coastal and oceanic bloom-dominating diatom *Skeletonema costatum* (SKE) were fed to *Calanus helgolandicus*, a dominant copepod in the temperate Atlantic. Controls were run with a small dinoflagellate, *Prorocentrum minimum* (PRO), typical of the bloom aftermath stage. We also followed the development of nauplii spawned from wild females collected during the recurrent winter/spring diatom bloom in the North Adriatic Sea¹¹, where *C. helgolandicus* individuals, overwintering at depth, rise to the surface and start first-feeding from late winter to early summer¹³.

C. helgolandicus females were conditioned with PRO for 24 h and then were either kept on this diet (control) or switched to a diet of SKE. Eggs were collected after three, five and seven days of feeding and allowed to hatch. Offspring were then raised on either PRO or SKE so that combinations of diets were tested: those in which both mothers and progeny always fed on either SKE or PRO, and those in which mothers received either PRO or SKE, and neonates were switched to the other diet.

Maternal and neonatal diet of SKE (SKE/SKE) resulted in 100% larval mortality, but the larval stage reached by the offspring

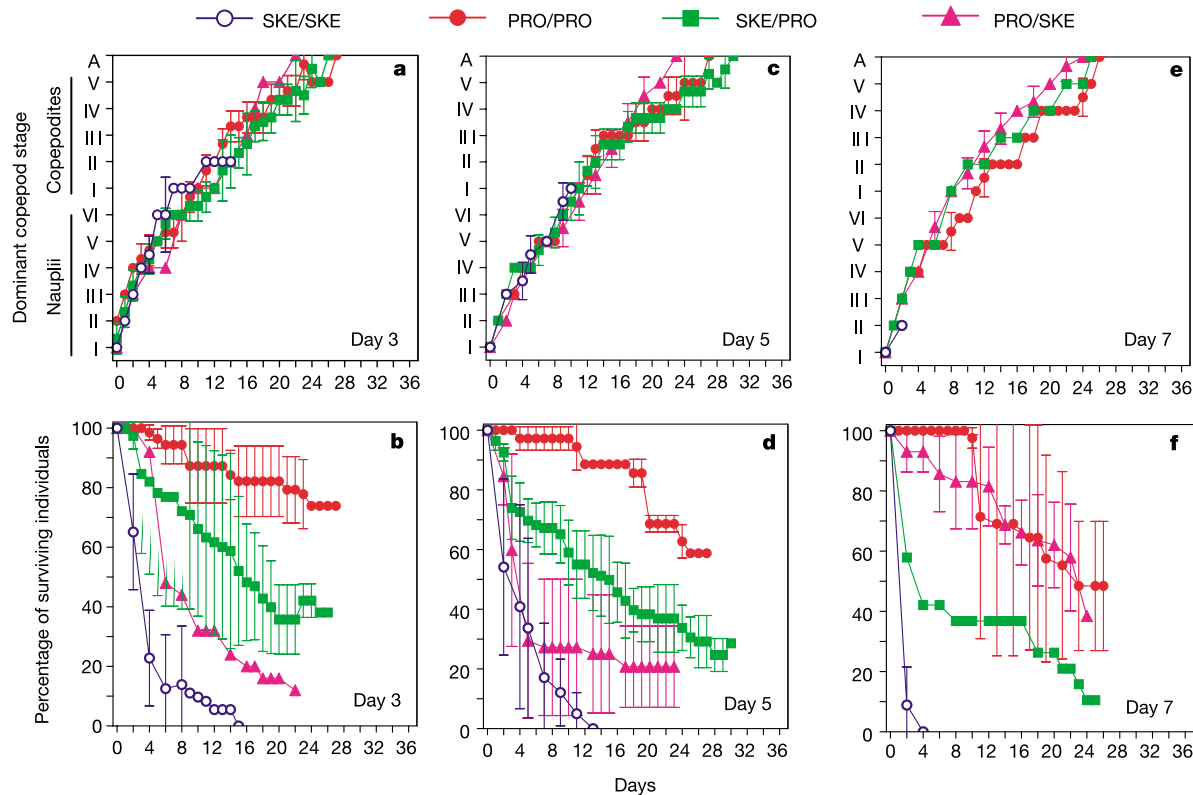


Figure 1 Effects of combinations of maternal and neonate diets on development rates (a, c, e) and percentage survivorship (b, d, f) of *C. helgolandicus* larvae. SKE/SKE and PRO/PRO are females and their larvae fed the diatom *S. costatum* (SKE) and the control dinoflagellate *P. minimum* (PRO), respectively. SKE/PRO indicates that females were fed SKE and their progeny fed PRO; the opposite is true for PRO/SKE. Experiments were run

with eggs collected from females fed either of the diets for three (a, b), five (c, d) or seven (e–f) days. Values are mean $\pm$ s.d. for one to three replicates of 20 newly hatched nauplii. Bars for development curves are smaller than for percentage survivorship because of greater individual variability in death rates compared with growth rates for the same group of animals.

depended on the days of exposure of the mother to the diatom diet (Fig. 1a–f). After three, five and seven days of feeding on SKE, development was arrested at copepodite stages CII, CI, and naupliar stage NII, respectively. Percentage survivorship was intermediate for the SKE/PRO and PRO/SKE diets and highest for PRO/PRO, although growth rates (reflected in development times to adulthood) were similar in all cases. Paired *t*-tests showed significant differences in daily survivorship but not in development rate (see Supplementary Information 1). The PRO diet resulted in decreasing survivorship with time (from 80% to 50% after three and seven days, respectively), suggesting a lack of some unknown essential nutrient(s) or other factor(s).

Nauplii collected from wild *C. helgolandicus* females during a diatom bloom from February to May 2003, dominated by SKE, *Stephanodiscus* sp. and *Chaetoceros* spp., did not survive to adulthood when reared on a diet of SKE (Fig. 2). Individuals died at or before the CII stage, except on the 13 and 27 March, when some nauplii (<20%) reached adulthood. Survivorship improved on the control diet PRO, but few reached adulthood except after late April.

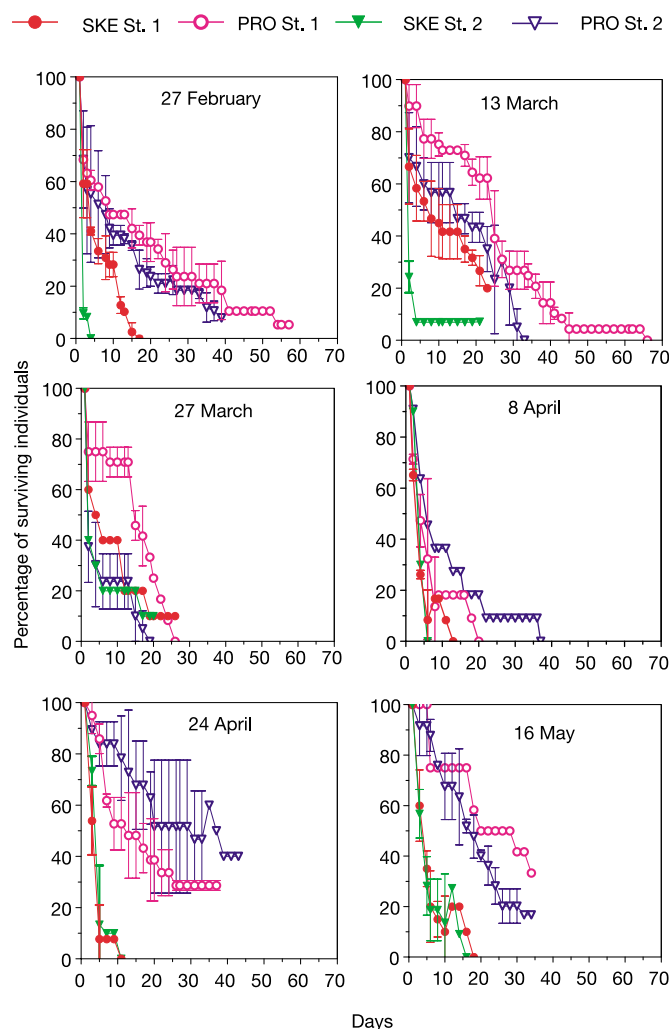


Figure 2 Percentage survivorship of *C. helgolandicus* nauplii spawned from wild females during the winter/spring bloom of 2003 in the North Adriatic Sea, and successively reared on the diatom *S. costatum* (SKE) or the dinoflagellate *P. minimum* (PRO). Values are mean $\pm$ s.d. for one to three replicates of 20 newly hatched nauplii. Experiments were run as in Fig. 1, on six sampling occasions and two different stations (St. 1 and St. 2). Adulthood was reached on only two occasions with SKE, with <20% survivorship, and on only three occasions with PRO, with highest survivorship (<40%) after mid-April.

Percentage egg viability during the bloom ranged from 16.8 ± 7.1 to 39.7 ± 7.5 (mean $\pm$ s.d.), similar to earlier results¹¹. Thus, maternal diatom diets in the field negatively affect not only egg hatching success, but also the development of nauplii that do hatch.

Mothers fed on SKE produced nauplii with strong teratogenic (congenital) birth defects. First described for *C. helgolandicus* reared on a diatom diet (*Phaeodactylum tricornutum*)¹⁴, the new experiments show that the reduction in egg viability (Fig. 3a) is accompanied by a concomitant rise in the number of hatched teratogenic nauplii (Fig. 3b). By day six, 45–65% of the hatched nauplii were malformed. Abnormal nauplii had asymmetrical bodies and malformed or reduced numbers of appendages (Fig. 3c, e, g). Some died within a few hours after hatching, whereas others died about one day later, at naupliar stage NII, because they were unable to swim or feed properly. Cell death through apoptosis had occurred in many tissues of the body, especially in appendages with strong structural malformations, and after nine days most nauplii were strongly deformed (Fig. 3d, f, h).

A diet of benthic diatoms can cause similar anatomical malformations in larvae of polychaetes and echinoderms¹⁵. The compounds responsible for these effects are short-chain $\alpha,\beta,\gamma,\delta$ -unsaturated aldehydes that arrest embryonic development in copepods and sea urchins, and have antiproliferative and apoptotic effects on human carcinoma cells¹¹. To test whether these aldehydes also arrested larval development in copepods, mothers were fed a culture of PRO to which concentrations of 0.5, 1.0 and $1.5 \mu\text{g ml}^{-1}$ of the diatom-derived aldehyde 2-*trans*, 4-*trans* decadienal (DD) were added. Nauplii were then successively raised on SKE (Fig. 4). DD was chosen as a model aldehyde because its toxic properties are currently being tested on various animal cells^{16,17,18}. The toxicity of DD and similar aldehydes more common in SKE¹⁹ are all related to a reactive Michael acceptor structural element that is characteristic of different fatty-acid-derived diatom metabolites^{3,20,21}. Because copepods do not drink²², and the chitinous exoskeleton is impervious to aldehydes^{18,23}, DD can enter the animals only by adsorption on ingested PRO cells.

The results obtained at $1.5 \mu\text{g ml}^{-1}$ were similar to those of feeding experiments (Fig. 1, SKE/SKE after three and five days), with nauplii dying at copepodite stage CI. At lower concentrations, some nauplii reached adulthood, but mortality was extremely high during the entire developmental phase. We calculated that females had ingested about 100 pg decadienal daily at $1.5 \mu\text{g ml}^{-1}$, which is even below the range animals will encounter in the field, considering aldehyde production of about 30 fg per cell in wounded SKE (T.W., unpublished results) and ingestion rates of about 1,000 diatom cells per hour²⁴.

Our findings bring new insights to the century-long debate on the role of the spring diatom bloom in pelagic food chains leading to commercial fish stocks. Larval growth rates of copepods on pure diatom diets indicate sufficient food quality with no need for supplementation. Besides, a feeding experiment in which two strains of the cosmopolitan, bloom-forming diatom *Thalassiosira rotula*, one lacking the fatty-acid-cleaving enzyme, were fed to copepods clearly showed that only aldehyde production impairs recruitment²⁰. A similar finding has been reported for an aldehyde-free strain of SKE fed to benthic animals²⁵. Because reproductive success depends on the amount of aldehydes ingested, individual grazers with the least predilection for aldehyde-producing diatoms will be selected for. This would explain the hatching variability observed in blooms from different areas¹². Indeed, copepods prefer motile protists (ciliates, dinoflagellates) over diatoms²⁶, so a mixed diet also serves to dilute the toxin²⁴. Selective copepod feeding on protistan grazers of the diatoms further facilitates biomass build-up of blooms. The ungrazed biomass eventually sinks out, with significant consequences for ocean ecology and biogeochemistry.

In contrast to benthic and terrestrial realms, the study of grazer defences in the plankton is in its infancy²⁷. Diatom frustules provide

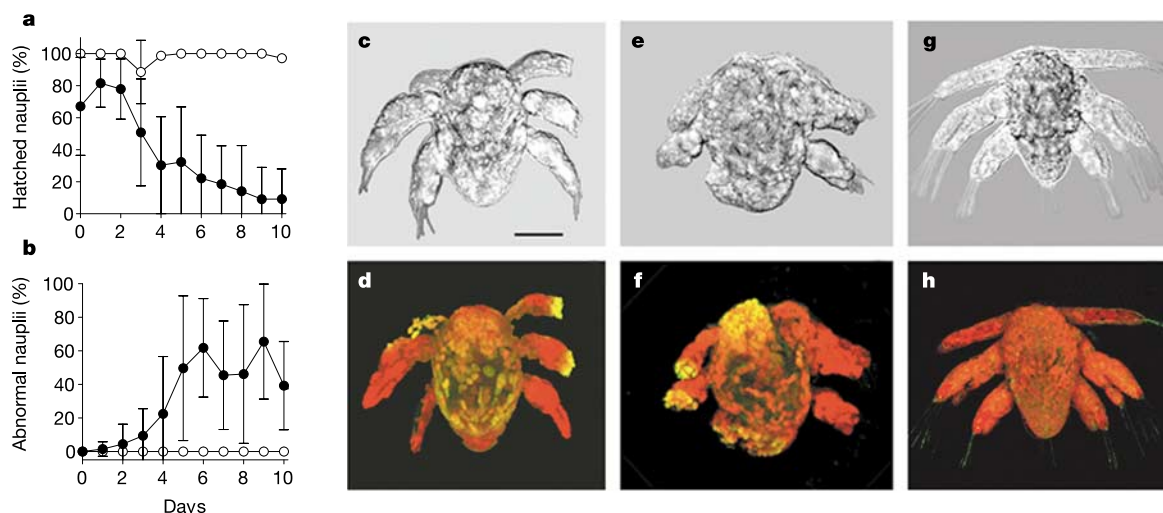


Figure 3 Effects of diet on *C. helgolandicus* offspring fitness. **a**, After ten days of feeding, the viability of eggs spawned by *C. helgolandicus* females fed the diatom *S. costatum* SKE (filled circles) dropped to <20% compared with >95% with the control dinoflagellate *P. minimum* PRO (open circles). **b**, After five days of feeding on SKE, 45–65% of the hatched nauplii were abnormal. **c**, **d**, Such nauplii had deformed limbs that were positive

for TUNEL staining (yellow, **d**) specific for apoptosis. **e**, **f**, After nine days of feeding on SKE, the degree of teratogenesis increased and nauplii were strongly deformed. **g**, **h**, Nauplii generated from females fed the control PRO diet appeared normal and stained negatively with TUNEL (**h**), indicating that nuclei were not apoptotic. Scale bar, 90 µm.

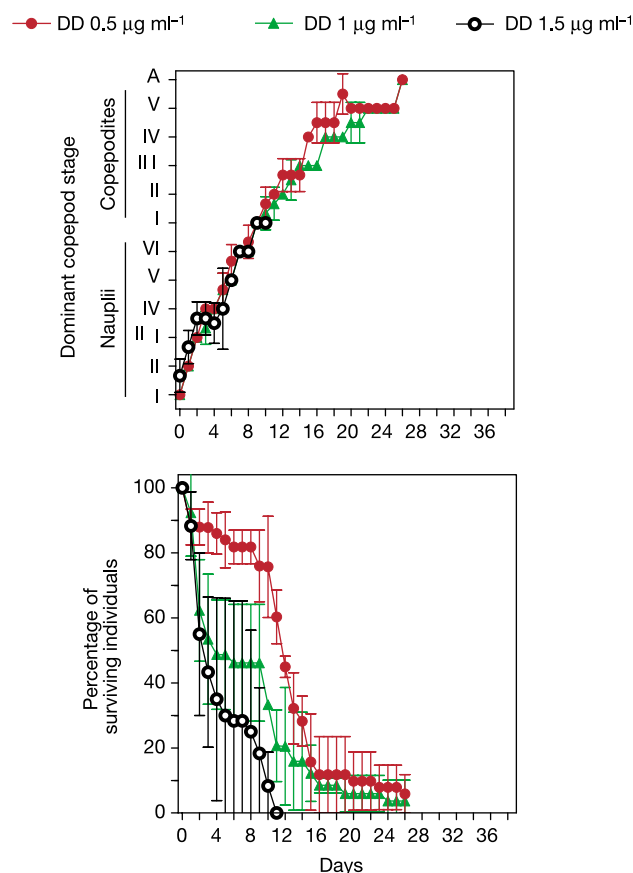


Figure 4 The effects of the aldehyde DD on development and percentage survivorship of *C. helgolandicus*. Females were fed the control dinoflagellate *P. minimum* (PRO) for three days, but with the addition of 0.5, 1.0 and 1.5 µg ml⁻¹ of decadienal. Nauplii were then raised on the diatom *S. costatum* (SKE). At lower concentrations, nauplii reached adulthood but survivorship was low. At the highest concentration, none of the nauplii reached adulthood and survivorship was only to copepodite stage Cl.

mechanical protection against some classes of grazers²⁸, but the effect of chemical defences based on teratogens will be more subtle. Apparently, the success of bloom-forming species lies in pairing metabolically ‘cheap’ mechanical and chemical defences with fast growth rates. Our findings provide a plausible mechanism for the apparent poor timing between bloom development and the arrival of the bulk of the overwintering copepod stock. Because not all bloom-forming species produce aldehydes, competition between defended and innocuous species could explain the complex oscillations²⁹ between diatom and copepod populations observed over decadal scales⁷. An evolutionary arms race operating in the plankton³⁰ could well be a major force shaping the annually recurring, yet unexplained patterns of plankton species succession. □

Methods

Rearing experiments

Female *C. helgolandicus* copepods, collected offshore of Roscoff (France) and Chioggia (Italy), were pre-conditioned for 24 h in crystallizing dishes filled with 100 ml of 0.22 µm filtered sea water enriched with the dinoflagellate *P. minimum* (PRO) at a final concentration of 8.6×10^3 cells per ml, and kept at 17°C and on a 12 h light/12 h dark cycle. Females were then divided into two groups and transferred to new containers with either PRO (control) at the same concentration, or the diatom *S. costatum* (SKE) at 7.3×10^4 cells per ml so that final algal concentrations were equivalent in terms of carbon (PRO = 177.1 pg per cell and SKE = 20.7 pg per cell). Methods for cultivation of algae are reported elsewhere²⁴. Females were transferred to new culture media daily. After three, five and seven days of feeding, one to three replicates of 20 N1, from females fed either PRO or SKE, were placed in 300-ml jars with 0.22 µm filtered sea water containing, in turn, either PRO or SKE, so that four combinations of diets were tested: those in which females and offspring both received PRO (control) or both received SKE; and those in which females received either PRO or SKE, and their offspring the other diet. Juveniles were collected daily, counted and checked under the microscope to determine the stage of development, and transferred to jars with fresh medium. A tally was kept of the most abundant larval stage and percentage of surviving individuals, to calculate development and survival curves. Rearing experiments were also conducted on nauplii collected from wild females during the winter/spring diatom bloom of 2003 in the North Adriatic Sea. Copepods were collected at two stations representative of the coastal (station 2), and offshore (station 1) areas, separated by a haline front⁶. Immediately hatched N1 were collected and transferred to SKE and PRO diets; developmental stage was determined as above. Rearing experiments testing the effects of DD (Sigma) were conducted using the same procedures as above. DD was dissolved in methanol to obtain a stock solution of 0.3 mg ml⁻¹. Females were fed PRO cultures to which 0.5, 1.0 and 1.5 µg ml⁻¹ of DD were added for three days; the methanol solvent added had no negative effect on copepods or their eggs, up to 1% in the final solution. Nauplii collected on day four were then reared on a diet of SKE.

Determining the amount of DD ingested by copepods

Samples of 40 ml of PRO (8.0×10^4 cells per ml) were incubated in sea water with $1.5 \mu\text{g ml}^{-1}$ of DD. After different incubation times, the cells were filtered onto GF/C filter (21 mm, Whatman) under reduced pressure until dry. The filter was rinsed with 1 ml 25 mM pentafluorobenzylhydroxylamine in 100 mM Tris-HCl, pH 7.0. After addition of benzaldehyde (5 μl of a 1 mM solution in methanol) as an internal standard, the cell suspension was sonicated in an ice bath with pulses of a B. Braun 1000 l Sonicator for 1 min. Afterwards, the sample was incubated at room temperature for 30 min. Extraction was performed with hexane³¹. Each treatment was replicated three times. Triplicate controls, consisting of sea water and DD without PRO, were conducted to determine the filter adsorption of DD. Detection was performed with gas chromatography/mass spectrometry (GC/MS) (GC Q; equipped with a 30 m RTX-200 column, 0.25 mm internal diameter, 0.25 μm film thickness). The analyses were performed by negative ion chemical ionization electron-capture mass spectrometry with methane as the reagent gas. For quantification of DD, the ion at m/z 327 $[\text{M}-\text{HF}]^-$ was chosen. A calibration curve shows linearity ($r^2 > 0.98$) in the measurement range (see Supplementary Information 2).

Assessment of Apoptosis

Egg hatching success and number of teratogenic nauplii were also monitored daily with the two diets. Procedures to determine egg-hatching success are described elsewhere²⁴. Apoptosis in teratogenic nauplii was verified using TdT-mediated dUTP nick end labelling (TUNEL) (Roche Diagnostics). *C. helgolandicus* nauplii were fixed overnight in 4% paraformaldehyde and 0.2 M NaCl in PBS, pH 7.4, rinsed in PBS, and frozen in liquid nitrogen to fracture the carapace. Samples were incubated for 24 h in 1 unit ml^{-1} chitinase enzyme (Sigma) at 25 °C, and rendered permeable according to the TUNEL manufacturer's instructions. They were then incubated for 90 min at 37 °C in TUNEL reaction mix and for 30 min in 0.5 $\mu\text{g ml}^{-1}$ propidium iodide at room temperature. Nauplii were observed with a confocal laser-scanning microscope, Zeiss LSM-410, in which TUNEL-positive areas appear yellow because of the superimposition of the green fluorescence of TUNEL and the red fluorescence of propidium iodide. Complementary tests with a mammalian cell line (A1 mes c-myc cells), generated from mouse mesencephalon primary cultures, suggest that DD is potentially a neutral compound for somatic but not for embryonic development, affecting undifferentiated proliferating rather than differentiated non-proliferating cells (see Supplementary Information 3).

Received 22 October 2003; accepted 30 March 2004; doi:10.1038/nature02526.

1. Smetacek, V. Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance. *Mar. Biol.* **84**, 239–251 (1985).
2. Cushing, D. H. *Marine Ecology and Fisheries* (Cambridge Univ. Press, Cambridge, 1975).
3. Pohnert, G. Wound-activated chemical defence in unicellular planktonic algae. *Angew. Chem. Int. Edn.* **39**, 4352–4354 (2000).
4. Ianora, A., Poulet, S. A. & Miralto, A. The effects of diatoms on copepod reproduction: a review. *Phycologia* **42**, 351–363 (2003).
5. Verity, P. G. & Smetacek, V. Organism life cycles, predation, and structure of pelagic ecosystems. *Mar. Ecol. Prog. Ser.* **130**, 277–293 (1996).
6. Miralto, A. *et al.* Inhibition of population growth in the copepods *Acartia clausi* and *Calanus helgolandicus* during diatom bloom. *Mar. Ecol. Prog. Ser.* **254**, 253–268 (2003).
7. Colebrook, J. M. Continuous plankton records: seasonal variation in the distribution and abundance of plankton in the North Atlantic Ocean and the North Sea. *J. Plankton Res.* **4**, 435–462 (1982).
8. Mauchline, J. *The Biology of Calanoid Copepods* (Academic, London, 1998).
9. Smetacek, V. The ocean's veil. *Nature* **419**, 565 (2002).
10. Ban, S. *et al.* The paradox of diatom copepod interactions. *Mar. Ecol. Prog. Ser.* **157**, 287–293 (1997).
11. Miralto, A. *et al.* The insidious effect of diatoms on copepod reproduction. *Nature* **402**, 173–176 (1999).
12. Irigoien, X. *et al.* Copepod hatching success in marine ecosystems with high diatom concentrations. *Nature* **419**, 387–389 (2002).
13. Weikert, H., Koppelman, R. & Wiegatz, S. Evidence of episodic changes in deep-sea mesozooplankton abundance and composition in the Levantine Sea (Eastern Mediterranean). *J. Mar. Syst.* **30**, 221–239 (2001).
14. Poulet, S. A. *et al.* Reproductive response of *Calanus helgolandicus*. I. Abnormal embryonic and naupliar development. *Mar. Ecol. Prog. Ser.* **129**, 85–95 (1995).
15. Caldwell, G. S., Olive, P. J. W. & Bentley, M. G. Inhibition of embryonic development and fertilization in broadcast spawning marine invertebrates by water soluble diatom extracts and the diatom toxin 2-trans, 4-trans decadienal. *Aquat. Toxicol.* **60**, 123–137 (2002).
16. Caldwell, G. S., Bentley, M. G. & Olive, P. J. W. The use of a brine shrimp (*Artemia salina*) bioassay to assess the toxicity of diatom extracts and short chain aldehydes. *Toxicol.* **42**, 301–306 (2003).
17. Tosti, E. *et al.* Bioactive aldehydes from diatoms block the fertilization current in ascidian oocytes. *Mol. Reprod. Dev.* **66**, 72–80 (2003).
18. Romano, G. *et al.* A marine diatom-derived aldehyde induces apoptosis in copepod and sea urchin embryos. *J. Exp. Biol.* **206**, 3487–3494 (2003).
19. d'Ippolito, G. *et al.* New birth-control aldehydes from the diatom *Skeletonema costatum*: characterization and biogenesis. *Tetrahedron Lett.* **43**, 6133–6136 (2002).
20. Pohnert, G. *et al.* Are volatile unsaturated aldehydes from diatoms the main line of chemical defence against copepods? *Mar. Ecol. Prog. Ser.* **245**, 33–45 (2002).
21. Adolph, S., Poulet, S. A. & Pohnert, G. Synthesis and biological activity of $\alpha,\beta,\gamma,\delta$ -unsaturated aldehydes from diatoms. *Tetrahedron* **59**, 3003–3008 (2003).
22. Tester, P. A. & Turner, J. T. Why is *Acartia tonsa* restricted to estuarine habitats? *Bull. Plankton Soc. Jpn* (Spec. Iss.) 603–611 (1991).
23. Takai, M. *et al.* Structure–property relationship of α - and β -chitin. *ACS Symp. Ser.* **489**, 38–52 (1992).
24. Turner, J. T. *et al.* Decoupling of copepod grazing rates, fecundity and egg-hatching success on mixed and alternating diatom and dinoflagellate diets. *Mar. Ecol. Prog. Ser.* **220**, 187–199 (2001).
25. Caldwell, G. S. *Diatom Mediated Disruption of Invertebrate Reproduction and Development*. PhD thesis, Univ. Newcastle-upon-Tyne (2004).
26. Kleppel, G. S. On the diets of calanoid copepods. *Mar. Ecol. Prog. Ser.* **99**, 183–195 (1993).

27. Wolfe, G. V. The chemical defense ecology of marine unicellular plankton: constraints, mechanisms, and impacts. *Biol. Bull.* **198**, 225–244 (2000).
28. Hamm, C. E. *et al.* Diatom cells are mechanically protected by their strong, lightweight, silica shells. *Nature* **421**, 841–843 (2003).
29. Yoshida, T. *et al.* Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* **424**, 303–306 (2003).
30. Smetacek, V. A watery arms race. *Nature* **411**, 745 (2002).
31. Luo, X. P. *et al.* Determination of aldehydes and other lipid-peroxidation products in biological samples by gas-chromatography mass-spectrometry. *Anal. Biochem.* **228**, 294–298 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Esposito for the preparation of algal cultures and assistance in feeding and rearing experiments. Thanks are also due to F. Palumbo, M. Perna, M. Di Pinto, J. M. Roualec and P. Quemener for technical assistance at sea. The authors acknowledge the financial contribution of their respective Institutes, and G.P. and T.W. also acknowledge that of the Deutsche Forschungsgemeinschaft. This paper represents a contribution towards the aims of MARBEF. MARBEF is an EU Network of excellence on Marine Biodiversity and Ecosystem Functioning under EU-Framework Programme 6.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.I. (ianora@szn.it).

Predator diversity dampens trophic cascades

Deborah L. Finke & Robert F. Denno

Department of Entomology, University of Maryland, College Park, Maryland 20742, USA

Food web complexity is thought to weaken the strength of terrestrial trophic cascades^{1–3} in which strong impacts of natural enemies on herbivores cascade to influence primary production indirectly⁴. Predator diversity can enhance food web complexity because predators may feed on each other and on shared prey^{5–7}. In such cases, theory suggests that the impact of predation on herbivores relaxes and cascading effects on basal resources are dampened^{8,9}. Despite this view, no empirical studies have explicitly investigated the role of predator diversity in mediating primary productivity in a natural terrestrial system^{10,11}. Here we compare, in a coastal marsh community, impacts of arthropod predators on herbivores and plant productivity between a simple food web with a single predator species and a complex food web with a diverse predator assemblage. We show that enhancing predator diversity dampens enemy effects on herbivores and weakens trophic cascades. Consequently, changes in diversity at higher trophic levels can significantly alter ecosystem function in natural systems.

Studies investigating the impact of biodiversity on ecosystem functions such as primary production have become widespread as a result of concern over the rapid rate of species extinctions^{10,12,13}. However, many studies in this area have focused specifically on the role of producer diversity, whereas the consequences of biodiversity loss at higher trophic levels have been often ignored^{11,13,14}, despite evidence that top trophic levels can be more susceptible to extinction than their basal resources^{11,15,16}. Studies that do incorporate trophic interactions into investigations of the link between biodiversity and ecosystem function have manipulated only the diversity of herbivores or filter-feeders^{17–19}, the diversity of consumers at several trophic levels simultaneously^{12,20,21}, or the overall presence or absence of predators^{2,22–24}. Few studies have independently manipulated predator diversity^{25,26} and none have done so in a natural terrestrial community. This study specifically examines the

importance of predator diversity for maintaining ecosystem function, and it does so in a native terrestrial salt marsh community that is vulnerable to human impacts.

We investigated the consequences of the loss of predator diversity for the occurrence of trophic cascades and its impact on primary productivity by using a natural assemblage of arthropods inhabiting the *Spartina* cordgrass-dominated salt marshes along the Atlantic coast of North America. Phloem-feeding *Prokelisia* planthoppers, the most abundant herbivores on the marsh, are consumed by a diversity of invertebrate predators including the hunting spiders *Pardosa littoralis* and *Hogna modesta*, the web-building spider *Grammonota trivittata* and the mirid bug *Tytthus vagus*^{27–29} (Fig. 1). Because *Grammonota*, *Tytthus* and *Pardosa* are susceptible to intra-guild predation^{27–29}—in which predators at the same trophic level feed on each other—the opportunity exists for antagonistic interactions among predators with cascading consequences for primary production.

In the context of this coastal marsh community, we constructed replicated food webs with various levels of predator species diversity (0, 1 or multiple predators) and measured the resulting impacts on herbivore population size and primary production. This study was conducted concurrently in the controlled setting of greenhouse mesocosms and under real-world conditions with the use of field enclosures at a marsh in Ocean County, New Jersey, USA. The species richness component of predator diversity was manipulated to create four food-web complexity treatments: (1) *Spartina* plants only, (2) *Spartina* plants and *Prokelisia* herbivores with no predators present, (3) *Spartina* plants, *Prokelisia* herbivores and a low-predator-diversity treatment (*Tytthus* only), and (4) *Spartina* plants, *Prokelisia* herbivores and a high-diversity predator assemblage (*Tytthus*, *Grammonota* and *Pardosa* in both mesocosms and field enclosures, and also with *Hogna* in mesocosms only). We manipulated predator diversity by using an additive treatment design to hold intraspecific interactions among *Tytthus* constant across levels of diversity³⁰ and to provide densities of predators that were equivalent to those found in the field (*Tytthus* 250 m⁻², *Grammonota* 250 m⁻², *Pardosa* 125 m⁻² and *Hogna* 25 m⁻²). Because we did not include treatments containing each predator individually, this treatment design does not allow a test of the null hypothesis of additive predator effects. However, *Tytthus*, *Grammonota*, *Pardosa* and *Hogna* are known to reduce planthopper population sizes independently in comparison with no-predator

controls^{27–29}. Therefore, although no conclusions can be made about the nature of predator interactions if planthopper populations decrease when predator diversity is high, an increase in planthopper populations in the presence of the predator complex would indicate unequivocally that antagonistic interactions among predators occur. To determine the influence of these predator diversity treatments on the strength of top-down effects, planthopper population density and plant productivity were assessed at the end of the study. Aboveground biomass and the number of tillers produced (an indirect measure of biomass in the following year) were determined as a proxy for primary productivity. Results are the consequences of longer-term food-web dynamics because the experiment spanned more than two herbivore generations from July to October 2002.

In greenhouse mesocosms, a trophic cascade occurred in the simple-structured food web with low predator diversity. This trophic cascade was diminished in the complex food web with a diverse predator assemblage. Planthopper density in the presence of the single predator was markedly reduced in comparison with the density of planthoppers when no predators were present ($F_{3,27} = 25.67$, $P < 0.0001$; $t = 6.27$, $P < 0.0001$; Fig. 2a). However, when predator diversity was high, planthopper density was

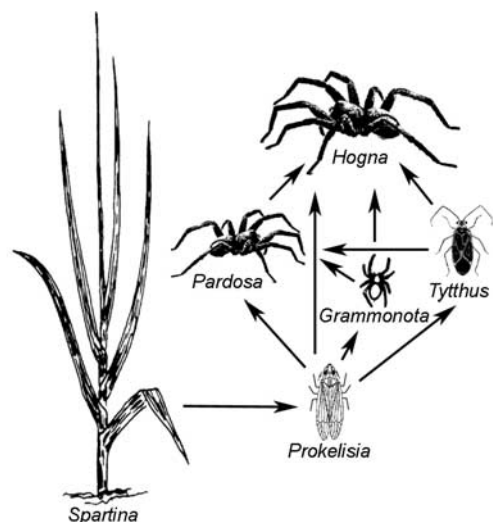


Figure 1 Component of salt marsh food web used in experimental design^{27–29}. Arrows indicate the flow of energy from the source to the consumer.

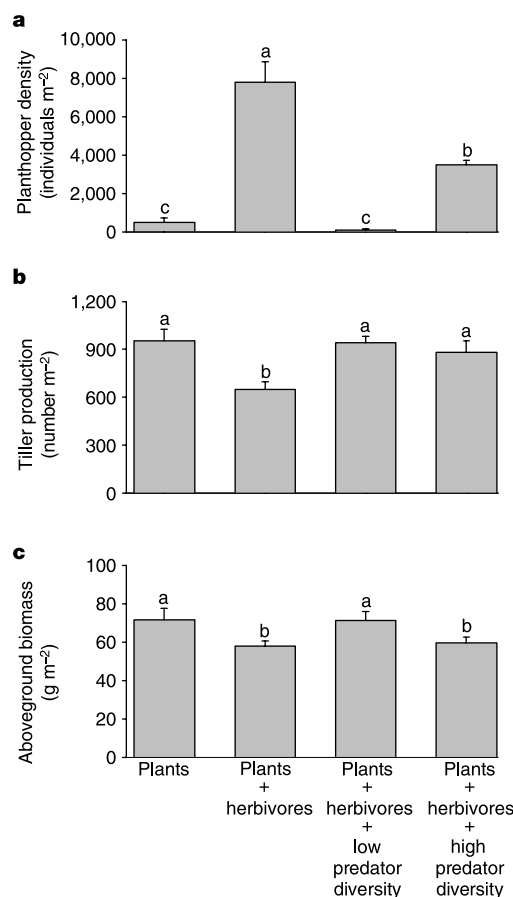


Figure 2 Effect of predator diversity on the occurrence of trophic cascades in greenhouse mesocosms. Means $\pm$ s.e.m. with different letters are significantly different ($P < 0.05$). **a**, Herbivore (*Prokelisia* planthopper) population size. Low predator diversity markedly reduces planthopper population size ($t = 6.27$, $P < 0.01$), but planthopper suppression is diminished when diversity is high ($t = 5.34$, $P < 0.01$). **b**, Number of tillers produced by *Spartina* cordgrass. In the absence of predators, planthoppers reduce the number of tillers ($t = 3.67$, $P < 0.01$). **c**, Aboveground biomass of *Spartina*. Low predator diversity enhances biomass relative to the high-predator-diversity treatment ($t = 2.44$, $P < 0.05$).

intermediate and greater than when predator diversity was low ($t = 5.34$, $P < 0.0001$; Fig. 2a). Predator effects on herbivore populations cascaded down to affect primary productivity, both tiller production ($F_{3,27} = 5.83$, $P < 0.01$; Fig. 2b) and aboveground live biomass ($F_{3,27} = 4.83$, $P < 0.01$; Fig. 2c). In the simple food web with a single predator species, the marked reduction in planthopper population size resulted in a trophic cascade, increasing both the number of tillers ($t = 3.52$, $P < 0.01$; Fig. 2b) and the aboveground biomass ($t = 2.83$, $P < 0.05$, Fig. 2c) in comparison with the predator-free herbivore treatment. In the high-predator-diversity treatment, the intermediate level of planthopper suppression was still sufficient to cascade down and increase the number of tillers ($t = 2.80$, $P < 0.05$; Fig. 2b) in comparison with the predator-free herbivore treatment. However, the intermediate control of the planthopper population by the predator complex did not cascade to affect aboveground biomass positively. Plant biomass in the complex food web with a diverse predator community was not different from that in the predator-free herbivore treatment ($t = 0.39$, $P > 0.05$; Fig. 2c). Thus, predator diversity precluded a trophic cascade on *Spartina* biomass because of the occurrence of intra-guild predation when predator diversity was high. Specifically, the population size of the *Tythus* mirid predator was much lower in the presence of other predators in the high-diversity treatment than

when alone in the low-predator-diversity treatment ($t = 4.07$, $P < 0.01$). This decline in density is attributed to intra-guild predation because spiders left small pellets of exsanguinated mirid exoskeletons after feeding²⁸. Therefore, when predator diversity was high, the occurrence of intra-guild predation resulted in an attenuation of enemy impacts on herbivores and dampened the strength of the trophic cascade on *Spartina* biomass.

Results of the field experiment were consistent with those from mesocosms. In the simple food web, predation by the single predator resulted in a trophic cascade. Planthopper density was reduced by *Tythus* predation ($F_{3,15} = 3.85$, $P < 0.05$; $t = 2.34$, $P < 0.05$; Fig. 3a), which increased the number of *Spartina* tillers ($F_{3,15} = 5.45$, $P < 0.01$; $t = 2.75$, $P < 0.05$; Fig. 3b) in comparison with the predator-free herbivore treatment. Treatment effects on aboveground biomass were not significant ($t = 1.10$, $P > 0.05$; Fig. 3c). In the complex food web with high predator diversity, the trophic cascade was dampened. There was no difference in planthopper densities when all predators were present and when no predators were present ($t = 0.29$, $P > 0.05$; Fig. 3a) and the density of tillers was also not different ($t = 0.92$, $P > 0.05$; Fig. 3b). Again, the dampening of the trophic cascade in the complex food web was due to the occurrence of intra-guild predation because the density of *Tythus* was significantly reduced in the presence of other predators ($t = 2.92$, $P < 0.01$). However, it is important to note that the strength of the cascade was weaker in the field than in mesocosms. This is probably the result of contamination of the field treatments by planthoppers (Fig. 3a) because of the latter's small size (3 mm) and high ambient density (about 11,000 individuals per m^2 during this study). This study therefore underscores the view that it might be more difficult to demonstrate trophic cascades in open systems than in closed systems.

Our results show that increasing the diversity of arthropod predators promotes intra-guild interactions among predators, diminishes enemy impacts on herbivores, and dampens cascading effects on basal resources. Therefore, given the widespread occurrence of intra-guild predators in natural systems⁶, a decline in predator species diversity might positively affect ecosystem function. A management conflict therefore arises because maximizing productivity, rather than preserving diversity, might be beneficial in certain contexts²⁴. For example, in agricultural systems the goal of biological-control programmes is to initiate trophic cascades by manipulating predator complexes to enhance crop yield, a circumstance that can arise when predator diversity is low or when antagonistic interactions among predators are minimal. By specifically examining the role of predator diversity, our study highlights how conservation biologists, whose goal is to maintain diversity, and biological-control practitioners, who seek to maximize productivity, can reach ultimately conflicting conclusions about the importance of biodiversity as it relates to ecosystem function. □

Methods

Greenhouse mesocosms

Greenhouse mesocosms consisted of ten field-collected *Spartina* culms transplanted into sand-filled pots (30 cm diameter, 0.04 m^2) and caged within a clear plastic cylinder (cellulose butyrate, 22 cm in diameter and 30 cm in height) sunk into the sand. Each mesocosm was covered by a screened lid (0.6 mm mesh, 85% light transmission). Forty mesocosms were placed into ten separate watering pools in groups of four (one replication of each treatment per pool) for a total of ten replications.

Field enclosures

Field enclosures were established in a *Spartina* meadow on an intertidal salt marsh in the Great Bay–Mullica River estuarine system in Tuckerton, Ocean County, New Jersey, USA. Circular enclosures (1.6 m^2 in area and 40 cm high) were constructed of PVC plastic sheeting covered with a screened lid (0.6 mm mesh, 85% light transmission) and sunk 10 cm into the marsh surface. To control for differences in elevation and grass height, the 24 enclosures were blocked in groups of four for a total of six replications.

Arthropod population density

Herbivore and predator densities were measured once at the end of each experiment. Densities within greenhouse mesocosms were determined by visually counting all

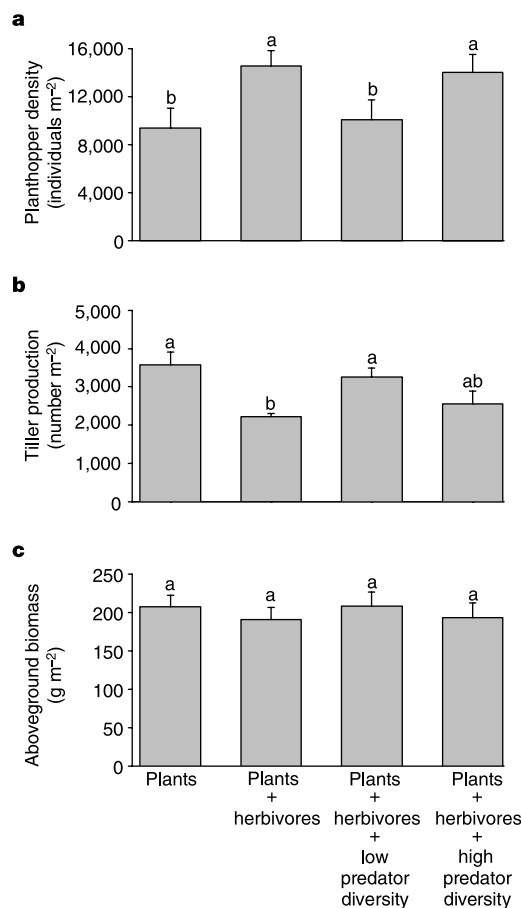


Figure 3 Effect of predator diversity on the occurrence of trophic cascades in field enclosures. Means $\pm$ s.e.m. with different letters are significantly different ($P < 0.05$). **a**, Herbivore (*Prokelisia* planthopper) population size. High predator diversity results in a population size no different from that when predators are absent ($t = 0.29$, $P > 0.05$). **b**, Number of tillers produced by *Spartina*. Low predator diversity enhances tiller number in comparison with the absence of predators ($t = 2.75$, $P < 0.05$). There is no difference in tiller production when predator diversity is high and when predators are absent ($t = 0.92$, $P > 0.05$). **c**, Aboveground biomass of *Spartina*. Predator diversity treatments did not impact biomass ($F = 0.72$, $P > 0.05$).

herbivores and predators. Densities within field enclosures were measured with an insect suction device. One sample consisted of eight 10-s placements of the sampling head on the marsh surface such that 0.8 m² of *Spartina* was subjected to suction.

Plant productivity

Two measures of plant productivity, aboveground biomass and the number of tillers produced (vegetative reproduction), were measured once at the end of the study. Aboveground biomass was determined for mesocosms by harvesting all live aboveground vegetation and for field enclosures by sampling all live aboveground biomass within a 0.047 m² wire frame. Vegetation was dried in an oven for 3 days at 55 °C and then weighed. The number of tillers produced was determined visually by counting all tillers in mesocosms and counting all tillers within the 0.047 m² sampling quadrat for the field enclosures.

Statistical analyses

The effects of the food web complexity treatments on final planthopper population size, the number of *Spartina* tillers and the aboveground biomass of *Spartina* were each analysed independently with mixed-model analyses of variance in which a block was modelled as a random source of variation. Subsequently, pairwise comparisons of treatment means were performed by using a *t*-test with a Bonferroni correction for multiple comparisons. Data were log-transformed when necessary to meet assumptions of normality and homogeneity of variances.

Received 16 February; accepted 6 April 2004; doi:10.1038/nature02554.

- Strong, D. R. Are trophic cascades all wet? Differentiation and donor-control in speciose ecosystems. *Ecology* **73**, 747–754 (1992).
- Halaj, J. & Wise, D. H. Terrestrial trophic cascades: How much do they trickle? *Am. Nat.* **157**, 262–281 (2001).
- Shurin, J. B. *et al.* A cross-ecosystem comparison of the strength of trophic cascades. *Ecol. Lett.* **5**, 785–791 (2002).
- Carpenter, S. R., Kitchell, J. F. & Hodgson, J. R. Cascading trophic interactions and lake productivity: fish predation and herbivory can regulate lake ecosystems. *Bioscience* **35**, 634–639 (1985).
- Crooks, K. R. & Soulé, M. E. Mesopredator release and avifaunal extinctions in a fragmented system. *Nature* **400**, 563–566 (1999).
- Polis, G. A., Myers, C. A. & Holt, R. D. The ecology and evolution of intraguild predation: potential competitors that eat each other. *Annu. Rev. Ecol. Syst.* **20**, 297–330 (1989).
- Rosenheim, J. A., Kaya, H. K., Ehler, L. E., Marois, J. J. & Jaffee, B. A. Intraguild predation among biological-control agents: theory and evidence. *Biol. Control* **5**, 303–335 (1995).
- Hart, D. R. Intraguild predation, invertebrate predators, and trophic cascades in lake food webs. *J. Theor. Biol.* **218**, 111–128 (2002).
- McCann, K. S., Hastings, A. & Huxel, G. R. Weak trophic interactions and the balance of nature. *Nature* **395**, 794–798 (1998).
- Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
- Duffy, J. E. Biodiversity loss, trophic skew and ecosystem functioning. *Ecol. Lett.* **6**, 680–687 (2003).
- Naem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
- Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845 (2001).
- Hooper, D. U. & Vitousek, P. M. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
- Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. & Torres, E. J. Fishing down marine food webs. *Science* **279**, 860–863 (1998).
- Petchey, O. L., McPherson, P. T., Casey, T. M. & Morin, P. J. Environmental warming alters food-web structure and ecosystem function. *Nature* **402**, 69–72 (1999).
- Cardinale, B. J., Palmer, M. A. & Collins, S. J. Species diversity enhances ecosystem functioning through interspecific facilitation. *Nature* **415**, 426–429 (2002).
- Norberg, J. Resource-niche complementarity and autotrophic compensation determines ecosystem-level responses to increased cladoceran species richness. *Oecologia* **122**, 264–272 (2000).
- Mulder, C. P. H., Koricheva, J., Huss-Danell, K., Höglberg, P. & Joshi, J. Insects affect relationships between plant species richness and ecosystem processes. *Ecol. Lett.* **2**, 237–246 (1999).
- McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
- Downing, A. L. & Leibold, M. A. Ecosystem consequences of species richness and composition in pond food webs. *Nature* **416**, 837–841 (2002).
- Gutiérrez, J. R., Meserve, P. L., Herrera, S., Contreras, L. C. & Jaksic, F. M. Effect of small mammals and vertebrate predators on vegetation in the Chilean semiarid zone. *Oecologia* **109**, 398–406 (1997).
- Fraser, L. H. & Grime, J. P. Top-down control and its effect on the biomass and composition of three grasses at high and low soil fertility. *Oecologia* **113**, 239–246 (1998).
- Schmitz, O. J. Top-predator control of plant biodiversity and productivity in an old-field ecosystem. *Ecol. Lett.* **6**, 156–163 (2003).
- Cardinale, B. J., Harvey, C. T., Gross, K. & Ives, A. R. Biodiversity and biocontrol: emergent impacts of a multiple-enemy assemblage on pest suppression and crop yield in an agroecosystem. *Ecol. Lett.* **6**, 857–865 (2003).
- Morin, P. J. & Lawler, S. P. Food web architecture and population dynamics: Theory and empirical evidence. *Annu. Rev. Ecol. Syst.* **26**, 505–529 (1995).
- Denno, R. F., Mitter, M. S., Langellotto, G. A., Gratton, C. & Finke, D. L. Interactions between a hunting spider and a web-builder: consequences of intraguild predation and cannibalism for prey suppression. *Ecol. Entomol.* (in the press).
- Finke, D. L. & Denno, R. F. Intraguild predation diminished in complex-structured vegetation: Implications for prey suppression. *Ecology* **83**, 643–652 (2002).
- Langellotto, G. *The Aggregation of Invertebrate Predators in Complex Habitats: Ecological Mechanisms and Practical Applications*. Thesis, Univ. Maryland (2002).
- Jolliffe, P. A. The replacement series. *J. Ecol.* **88**, 371–385 (2000).

Acknowledgements We thank B. Cardinale, W. Fagan, M. Palmer and O. Schmitz for their comments. This work was supported by grants from the National Science Foundation to R.E.D.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L.F. (dfinke@umd.edu).

Object-based attention determines dominance in binocular rivalry

Jude F. Mitchell¹, Gene R. Stoner² & John H. Reynolds¹

¹Systems Neurobiology Laboratory, and ²Vision Center Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA

A question of long-standing interest to philosophers, psychologists and neuroscientists is how the brain selects which signals enter consciousness^{1,2}. Binocular rivalry and attention both involve selection of visual stimuli, but affect perception quite differently. During binocular rivalry, awareness alternates between two different stimuli presented to the two eyes. In contrast, attending to one of two different stimuli impairs discrimination of the ignored stimulus, but without causing it to disappear from consciousness. Here we show that despite this difference, attention and rivalry rely on shared object-based selection mechanisms. We cued attention to one of two superimposed transparent surfaces and then deleted the image of one surface from each eye, resulting in rivalry. Observers usually reported seeing only the cued surface. They were also less accurate in judging unpredictable changes in the features of the uncued surface. Our design ensured that selection of the cued surface could not have resulted from spatial, ocular or feature-based mechanisms. Rather, attention was drawn to one surface, and this caused the other surface to be perceptually suppressed during rivalry. These results raise the question of how object representations compete during these two forms of perceptual selection, even as the features of those objects change unpredictably over time.

The relationship between attention and rivalry has been debated from the late nineteenth century^{1,2} to the present^{3,4}. The question of what is selected in attention and rivalry has also been disputed. It is well established that spatial locations can be selectively attended^{5,6}, but it is now recognized that objects can be selected as well^{7,8}. For rivalry, the debate has been whether competition is stimulus-based, eye-based or some combination of the two⁹. Using the paradigm illustrated in Fig. 1, we asked whether selection of an object by attention causes that object to be dominant during rivalry. Observers viewed two superimposed patterns of dots presented to both eyes at the start of each trial. The patterns rotated rigidly in opposite directions around a fixation point, yielding a percept of two superimposed transparent surfaces. After a period of dual rotation, one surface was briefly translated in one of eight directions, and the observer reported the perceived direction. Such brief translations are known to cue attention to the translated surface^{10–13}. Hence, we refer to the translated surface as the ‘cued surface’.

After translation, the image of the cued surface was removed from one eye and the image of the uncued surface was removed from the other eye (see Methods). Because the surfaces differed in rotation direction, this dichoptic presentation produced rivalry. To determine whether rivalry favoured the cued surface, we asked observers to report whether one surface was clearly dominant at the end of

dichoptic presentation, and if so, which surface. By varying the duration of dichoptic presentation from trial to trial, we traced the time course of dominance from 0 to 1,850 ms.

As illustrated in Fig. 2a, 150 ms after the switch to dichoptic presentation, the dominance of the cued surface was small but significant. This advantage was, however, quickly amplified. After 300 ms, rivalry was perceived on the majority of trials, and the cued surface was usually dominant. The cued surface remained dominant for at least 900 ms. The translation that cued attention was presented to both eyes. Nonetheless, the cued surface was dominant, regardless of whether it appeared in the right or the left eye during subsequent dichoptic viewing. Therefore, the dominance of the cued surface must have resulted from a selection mechanism that bypassed interocular competition.

In addition to circumventing eye-based mechanisms, our paradigm also ruled out spatial and feature-based selection. Spatial selection cannot account for our results, because the dominant and suppressed surfaces always occupied the same region of visual space. Nor can the dominance of the cued surface be explained by selection of an individual feature: it is known that observers can track the features of one of two superimposed stimuli as they change smoothly through feature space¹⁴. However, our study precludes this possibility because rotation followed translation abruptly. After translation, both directions of rotation were present, and only surface identity linked the cueing translation to the particular rotation direction later reported as dominant. Having ruled out ocular, spatial and feature-based selection, and having found that the identity of the cued surface predicted which rotation direction was dominant, we conclude that this selection is object-based.

Investigations of object-based attention have shown that performance in judging multiple visual features is better if the features belong to a single object, as opposed to multiple objects. This has been found to be true both when the two features are defined within different dimensions, such as colour and orientation^{7,14}, and when they are defined within the same dimension, such as visual motion^{10–13}. The latter set of experiments revealed that the second of two successive translations is poorly discriminated if the translations are of two surfaces, as opposed to when they are of the same surface.

In a second experiment, we adapted our paradigm to see whether these object-based effects would also be seen during rivalry. As in experiment 1, both surfaces rotated and then one surface briefly translated in one of eight directions. The observer reported perceived translation direction. One hundred and fifty milliseconds after this cueing translation, we again deleted the image of one surface from each eye, but instead of reporting which surface was then dominant, observers reported the direction of a second translation. We measured accuracy as a function of the time between the end of the cueing translation and the start of the second translation (the interstimulus interval).

To compare attentional selection during rivalrous and non-rivalrous viewing, we interleaved these rivalrous trials with non-rivalrous trials in which the stimuli never became dichoptic and were thus perceived as superimposed transparent surfaces throughout the trial. On half of these ‘transparency’ trials, both surfaces were presented to both eyes throughout the trial, a condition that we refer to as ‘binocular transparency’. On the other half of the transparency trials, both surfaces were deleted from one eye 150 ms after the first translation. This ‘monocular transparency’ condition controlled for any momentary disruption caused by deleting surfaces during the switch from normal binocular viewing to rivalrous viewing. But monocular transparency did not induce rivalry, as there was no competing stimulus in the other eye. Despite this transient event when switching to monocular transparency, there were no significant differences in the observers’ performance on monocular and binocular transparency trials (three-way analysis

of variance with interstimulus interval, viewing condition and surface as factors; see Supplementary Information). We therefore focus on monocular transparency, because it provides a more direct comparison to rivalry.

Observers were impaired in judging the second translation of the uncued surface during both rivalry (Fig. 2b) and monocular transparency (Fig. 2c), but with different time courses. In monocular transparency, the impairment was strongest 150 ms after the first translation. In contrast, the peak impairment during rivalry occurred 450 ms after the first translation. The impairment disappeared within 1,000 ms during transparency, but persisted in rivalry at the longest duration tested. The magnitude and time course of the impairment were significantly different in rivalry and monocular transparency, according to a three-way analysis of variance, with interstimulus interval, viewing condition and surface as factors ($P < 0.05$ for all two-way interactions; see Supplementary Information).

These double translation experiments isolated the contribution of neurons that mediate interocular competition from those that mediate object-based competition. Transparency and dichoptic

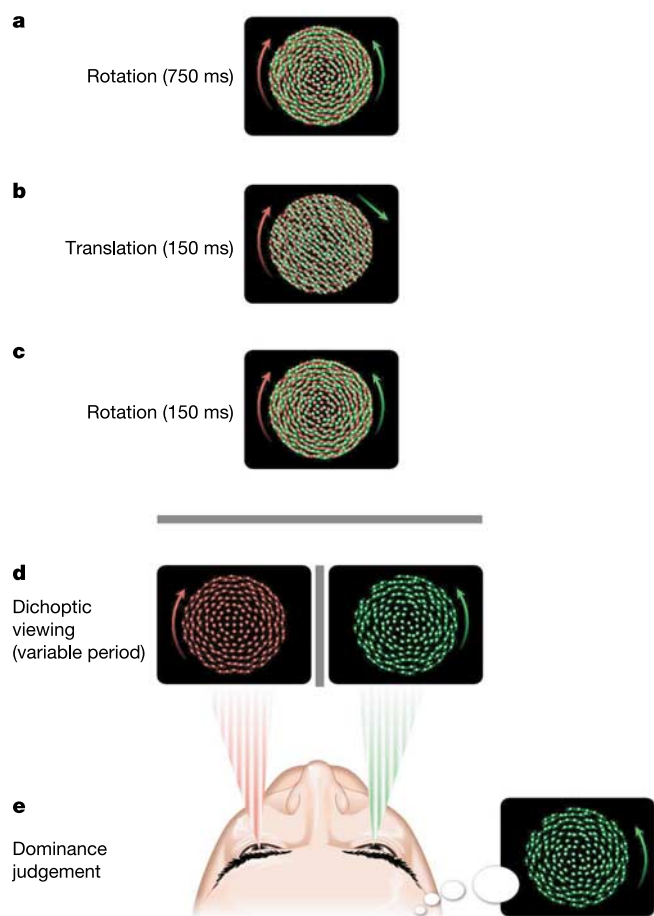


Figure 1 Dominance-judgement task. Panels are arranged from top to bottom according to the sequence of events that occurred on each trial in experiment 1. **a**, Two sets of dots rotated in opposite directions around a common point, yielding a percept of superimposed transparent surfaces viewed through an aperture. For illustration purposes, we show the surfaces with different colours (red and green), but they were actually the same colour. **b**, One of the surfaces, the green surface in this illustration, translated for 150 ms in one of eight directions, while the other surface continued to rotate. Subjects reported the direction of translation. **c**, Both surfaces resumed rotation for 150 ms. **d**, The image of one surface was removed from each eye, resulting in rivalry. Subjects judged which surface was dominant at the end of this variable-length period of rivalrous viewing. **e**, Observers usually perceived the previously translated surface as dominant.

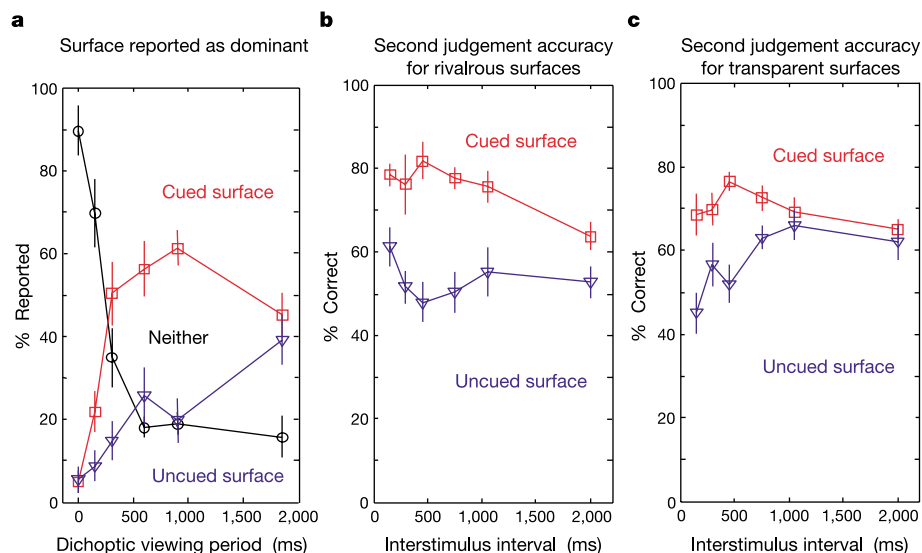


Figure 2 Data from dominance-judgement and double-translation tasks. **a**, Seven observers reported whether either surface was dominant at the end of dichoptic presentation and, if so, which surface was dominant. The mean percentage of trials on which the cued or uncued surface was reported to be dominant is shown in red and blue, respectively. The percentage of trials on which neither surface was clearly dominant appears in black. **b**, Mean accuracy in reporting the direction of the second translation averaged across trials in which the surfaces were presented dichoptically. **c**, Mean

accuracy in reporting the direction of the second translation when both surfaces were presented to one eye, and thus appeared transparent. Line colour in **b** and **c** indicates whether the cued (red line) or uncued (blue line) surface translated second. The interstimulus interval is the delay between the end of the cueing translation and start of the second translation (interstimulus interval = dichoptic viewing period + 150 ms). Error bars indicate standard errors of the mean across subjects.

viewing trials began identically, with one of the two surfaces cued by a sudden translation. Stimulus conditions were identical during cueing, so the same neurons logically must have been engaged, regardless of whether transparent or rivalrous viewing ensued. From this, we surmise that the same object-based mechanisms initiated selection for both transparency and rivalry trials. However, after the cueing phase, the sole difference between the rivalry and monocular transparency conditions was that the two deleted stimuli were removed from different eyes in one condition and from a single eye in the other. Thus, whereas dominance during rivalry and transparency must have been triggered by the same object-based mechanisms, the differences in the time courses of selection can only be due to neurons with eye-of-origin information. One intriguing difference is that the cueing effect was strongest at the shortest latency for transparency, but was initially disrupted during rivalry. This delay agrees with the relatively weak perceptual dominance found in experiment 1 at the shortest period of dichoptic viewing (Fig. 2a) and may reflect the time required for object-based mechanisms to influence interocular competition.

Our findings illuminate the long-standing debate about the relationship between rivalry and attention^{1,2}. Although some have concluded that attention has no influence on rivalry, recent studies suggest otherwise. Cueing spatial⁴ and feature-based¹⁵ attention during rivalry have been found to influence perceptual dominance. The present results show that when an object is cued during normal binocular viewing, it then dominates during subsequent rivalry. Besides revealing an unambiguous influence of attention, these results establish a connection between object-based attention and rivalry.

The present experiments also add to the literature showing that rivalry is not exclusively eye-based. Before our study, it was known that under certain conditions a stimulus can maintain its dominance, even when swapped between eyes^{16,17}. It was also known that complementary parts of two stimuli can be distributed across the eyes and fused into a coherent whole^{18–25}. Feature-based mechanisms can explain some of these earlier results, but other findings implicate more sophisticated, potentially object-based, mechanisms^{22,25}.

The present findings extend these earlier studies by showing that competition during rivalry can occur between representations of objects that change their attributes (here, directions of motion) unpredictably. This requires that the different motions be ‘bound’ into a coherent representation. Although the underlying mechanism is unknown and controversial^{26,27}, the binding of attributes is a hallmark of object-based representation^{26,27}.

Finally, these studies are consistent with the view that binocular rivalry, like other multi-stable phenomena, such as the Necker cube and the face–vase illusion, involves competition between high-level stimulus representations^{28–30}. Our results raise the possibility that selection of one of multiple possible interpretations of the visual scene (as with multi-stable percepts) and selection of one of several objects in the visual scene (during selective attention) engage common competitive mechanisms. □

Methods

Subjects viewed stimuli through a mirror stereoscope in a dark, quiet room. Sessions began with adjustment of the mirrors to allow binocular fusion of two nonius lines, displayed dichoptically on a Trinitron Multiscan TC running at 60 Hz. A bite bar stabilized the head while eye position was monitored using infrared tracking operating at 60 Hz (ISCAN). Trials terminated if fixation strayed from a one-degree square window. Reliable eye positions were obtained in five of seven subjects. The other subjects had eye shapes that precluded eye position monitoring, but showed similar cueing effects. The luminances of the left and right images were adjusted to compensate for ocular biases. Subjects discriminated a brief translation during rivalrous presentation of two, photometrically isoluminant (68 candelas per m²) dot patterns. If mean accuracy was better for translations presented to one eye, then luminance in that eye was reduced to equate performance. Both surfaces were the same colour, either red or green, chosen with equal probability. All experimental sessions consisted of 5 blocks of 100 trials, lasting 1.5–2 h with 5–10 min breaks between blocks. All judgements were indicated by key press.

Experiment 1: dominance-judgement task

Each trial began with fixation of a central grey point (0.25 × 0.25 degrees of visual arc) presented to both eyes. After the subject had pressed a key, two circular patterns of dots appeared, rotating in opposite directions around a central fixation point (see Fig. 1). These binocularly presented patterns appeared as two overlapping, rigidly moving, transparent surfaces. On each trial, one surface was selected at random to rotate clockwise and the other rotated anticlockwise. Each surface subtended 5.5 degrees of visual arc, was composed of 120 dots (averaging 5 dots per square degree of arc), and rotated at 50 degrees per second.

Trials began with 750 ms of dual rotation, after which one surface, chosen at random,

briefly (150 ms) translated in one of eight directions while the other surface continued rotating. Sixty per cent of the dots translated coherently while each of the remaining dots was randomly assigned to one of the other 7 directions, thus discouraging tracking of individual dots. The translation velocity was 1.5 degrees of visual arc per second. Observers reported translation direction at any time during the trial. Translation judgements were accurate (mean, 86.7% correct). Breaks of fixation, incorrect responses and correct responses were signalled by different sounds.

After translation, both surfaces rotated for 150 ms. The image of one surface, selected at random, was then deleted from one eye and the image of the other surface was deleted from the other eye. The surfaces then continued rotating for one of six dichoptic viewing periods (0, 150, 300, 600, 900 or 1,850 ms), after which both disappeared. Observers reported whether the clockwise, anticlockwise or neither surface had been dominant at the end of the trial. To confirm accuracy of these reports, we interleaved catch trials (20% of the total), in which either one surface appeared in both eyes or both surfaces appeared in both eyes. On most trials subjects correctly reported that neither surface was dominant when both were present (82.8% correct), and correctly identified the direction of rotation when a single surface was present (85.1% correct).

Experiment 2: double-translation task

The double-translation task was identical to the dominance-judgement task except that subjects judged a second translation (in one of eight directions) rather than dominance. The surface that underwent this second translation was selected at random on each trial. Translation was subsequently masked by 500 ms of dual rotation.

Three conditions were interleaved. In the rivalry condition, as in the dominance-judgement task, surfaces were presented dichoptically 150 ms after the cueing translation ended. In the monocular transparency condition, the images of both surfaces were deleted from one eye at this same time point. In the binocular transparency condition, both surfaces were presented throughout the trial.

We traced the time course of selection by varying the interstimulus interval (the period between the end of the cueing translation and the onset of the second translation). We used six interstimulus intervals (150, 300, 450, 750, 1,050 or 2,000 ms). The interstimulus interval was selected at random on each trial.

Subjects practised until they achieved 70% accuracy in judging the second translation of the cued surface under non-rivalrous conditions. Data from practice sessions were discarded. We then ran four experimental sessions (20 blocks in total). From session to session we adjusted the duration of the second translation so that the mean performance in judging the second translation of the cued surface remained approximately 70% for all conditions. The duration was varied from 33 to 100 ms in each condition.

Observers

Seven observers (five women and two men) participated in both of the main experiments. All had normal, or corrected to normal, vision. Five were naive. Ages ranged from 19 to 31 years.

Received 11 October 2003; accepted 15 April 2004; doi:10.1038/nature02584.

- James, W. *The Principles of Psychology* Vol. 1 (Henry Holt, New York, 1890).
- von Helmholtz, H. *Handbuch der physiologischen Optik* 3rd edn (Voss, Hamburg, 1909).
- Lack, L. *Selective Attention and the Control of Binocular Rivalry* (Mouton, The Hague, 1978).
- Ooi, T. L. & He, Z. J. Binocular rivalry and visual awareness: the role of attention. *Perception* **28**, 551–574 (1999).
- Posner, M. I. Orienting of attention. *Q. J. Exp. Psychol.* **32**, 3–25 (1980).
- Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
- Duncan, J. Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.* **113**, 501–517 (1984).
- Schoenfeld, M. A. *et al.* Dynamics of feature binding during object-selective attention. *Proc. Natl Acad. Sci. USA* **100**, 11806–11811 (2003).
- Blake, R. R. & Logothetis, N. K. Visual competition. *Nature Rev. Neurosci.* **3**, 13–23 (2002).
- Valdes-Sosa, M., Cobo, A. & Pinilla, T. Attention to object files defined by transparent motion. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 488–505 (2000).
- Pinilla, T., Cobo, A., Torres, K. & Valdes-Sosa, M. Attentional shifts between surfaces: effects on detection and early brain potentials. *Vision Res.* **41**, 1619–1630 (2001).
- Reynolds, J. H., Alborzian, S. & Stoner, G. R. Surface-based exogenous cueing triggers automatic competitive selection. *Vision Res.* **43**, 59–66 (2003).
- Mitchell, J. F., Stoner, G. R., Fallah, M. & Reynolds, J. H. Attentional selection of superimposed surfaces cannot be explained by modulation of the gain of color channels. *Vision Res.* **43**, 1323–1328 (2003).
- Blaser, E., Pylyshyn, Z. W. & Holcombe, A. O. Tracking an object through feature space. *Nature* **408**, 196–199 (2000).
- Sasaki, H. & Gyoba, J. Selective attention to stimulus features modulates interocular suppression. *Perception* **31**, 409–419 (2002).
- Logothetis, N. K., Leopold, D. A. & Sheinberg, D. L. What is rivaling during binocular rivalry? *Nature* **380**, 621–624 (1996).
- Lee, S. H. & Blake, R. Rival ideas about binocular rivalry. *Vision Res.* **39**, 1447–1454 (1999).
- Diaz-Caneja, E. Sur l'alternance binoculaire. *Ann. Oculist* October, 721–731 (1928).
- Whittle, P., Bloor, D. C. & Pocock, S. Some experiments on figural effects in binocular rivalry. *Percept. Psychophys.* **4**, 183–188 (1968).
- Kulikowski, J. J. Binocular chromatic rivalry and single vision. *Ophthalmol. Physiol. Opt.* **12**, 168–170 (1992).
- Kovacs, L., Papathomas, T. V., Yand, M. & Feher, A. When the brain changes its mind: interocular grouping during binocular rivalry. *Proc. Natl Acad. Sci. USA* **93**, 15508–15511 (1996).
- Alais, D. & Blake, R. R. Interactions between global motion and local binocular rivalry. *Vision Res.* **38**, 637–644 (1998).
- Alais, D. & Blake, R. R. Grouping visual features during binocular rivalry. *Vision Res.* **39**, 4341–4353 (1999).

- Sobel, K. V. & Blake, R. How context influences predominance during binocular rivalry. *Perception* **31**, 813–824 (2002).
- Ooi, T. L. & He, Z. J. A distributed interocular processing of binocular rivalry: psychophysical evidence. *Perception* **32**, 155–166 (2003).
- Castelo-Branco, M., Goebel, R., Neuenschwander, S. & Singer, W. Neural synchrony correlates with surface segregation rules. *Nature* **405**, 685–689 (2000).
- Thiele, A. & Stoner, G. Neural synchrony does not correlate with motion coherence in cortical area MT. *Nature* **421**, 366–370 (2003).
- Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549–553 (1996).
- Sheinberg, D. L. & Logothetis, N. K. The role of temporal cortical areas in perceptual organization. *Proc. Natl Acad. Sci. USA* **94**, 3408–3418 (1997).
- Leopold, D. A. & Logothetis, N. K. Multistable phenomena: changing views in perception. *Trends Cogn. Sci.* **3**, 254–264 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Crick, G. Horwitz, H. Jordan, B. Krekelberg, M. Richert and K. Sundberg for providing comments on the manuscript, and J. Reyes and C. Williams for testing subjects. J.M. was funded by an NIH Training Grant in Cognitive Neuroscience, G.S. was funded by an NEI grant and J.R. was funded by grants from NEI, The Sloan Foundation, and The McKnight Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R. (reynolds@salk.edu).

VEGF delivery with retrogradely transported lentivector prolongs survival in a mouse ALS model

Mimoun Azzouz¹, G. Scott Ralph¹, Erik Storkebaum², Lucy E. Walmsley^{1*}, Kyriacos A. Mitrophanous¹, Susan M. Kingsman¹, Peter Carmeliet² & Nicholas D. Mazarakis¹

¹Oxford BioMedica plc, The Oxford Science Park, Medawar Centre, Oxford OX4 4GA, UK

²Center for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotechnology, University of Leuven, Campus Gasthuisberg, Herestraat 49, B-3000 Leuven, Belgium

* Present address: Oxxon Pharmaccines, Florey House, 3 Robert Robinson Avenue, Oxford Science Park, Oxford, OX4 4GP, UK

Amyotrophic lateral sclerosis (ALS) causes adult-onset, progressive motor neuron degeneration in the brain and spinal cord, resulting in paralysis and death three to five years after onset in most patients¹. ALS is still incurable, in part because its complex aetiology remains insufficiently understood. Recent reports have indicated that reduced levels of vascular endothelial growth factor (VEGF), which is essential in angiogenesis and has also been implicated in neuroprotection^{2–4}, predispose mice and humans to ALS^{5,6}. However, the therapeutic potential of VEGF for the treatment of ALS has not previously been assessed. Here we report that a single injection of a VEGF-expressing lentiviral vector into various muscles delayed onset and slowed progression of ALS in mice engineered to overexpress the gene coding for the mutated G93A form of the superoxide dismutase-1 (SOD1^{G93A}) (refs 7–10), even when treatment was only initiated at the onset of paralysis. VEGF treatment increased the life expectancy of ALS mice by 30 per cent without causing toxic side effects, thereby achieving one of the most effective therapies reported in the field so far.

Because delivery of therapeutically attractive neuroprotective factors to motor neurons is a formidable challenge in ALS therapy, we recently developed a lentiviral gene transfer system to deliver

transgenes to neurons^{11,12}. To establish quantitatively the potential of these vectors—rabies-G pseudotyped equine infectious anaemia virus (EIAV)—to transfer genes to motor neurons in mice with neurodegeneration, we first evaluated their gene-transfer efficiency to motor neurons in SOD1^{G93A} mice, before and at onset of motor neuron degeneration, that is, at 21 and 90 days of age. Mutations in the SOD1 gene cause ALS in humans¹³ and because SOD1^{G93A} mice develop severe motor neuron degeneration closely resembling the human disease, they represent the best available animal model for ALS^{7,9,10}. Four weeks after the injection of high titre EIAV vector, carrying the LacZ reporter gene (EIAV–LacZ, $\sim 1 \times 10^9$ transducing units (TU) per ml), unilaterally into the gastrocnemius and facial muscles of 21-day-old SOD1^{G93A} mice, extensive reporter gene expression was observed in both the spinal cord and brainstem (Fig. 1a–c). We scored the motor neurons as being large neurons in the ventral horn with distinct nuclei and nucleoli, and determined that more than 700 motor neurons—that is, the majority of motor neurons^{14,15} innervating the gastrocnemius muscle—were transduced (Fig. 1a, d). Moreover, by staining for calcitonin gene related peptide (CGRP) to identify the spinal motor neurons more directly, we found that more than 60% of CGRP-positive motor neurons were transduced (Fig. 1e, f). Similar gene transfer studies in 90-day-old injected SOD1^{G93A} mice resulted in fewer transduced motor neurons (Fig. 1d). This does not reflect impaired gene transfer

efficiency in diseased ALS mice, but instead is attributable to the much smaller population of residual surviving motor neurons at disease onset¹⁶. Extensive reporter gene expression was also observed in facial (Fig. 1b, d) and motor trigeminal nuclei (Fig. 1c, d). These data show that EIAV-based lentiviral vectors pseudotyped with the rabies-G envelope glycoprotein are very efficient vectors for the remote delivery of transgenes to motor neurons after injection into muscle.

VEGF^{Δ/Δ} ‘knock-in’ mice, in which the hypoxia-response element sequence in VEGF promoter was deleted, have an impaired potential to upregulate VEGF levels in conditions of stress. These mice develop ALS-like neuropathology³, suggesting that motor neurons are particularly sensitive to reductions in the levels of VEGF. VEGF has also been reported to have favourable effects on ischaemic neuropathy in mice³. We hypothesized that VEGF gene transfer to motor neurons might slow down motor neuron degeneration in SOD1^{G93A} mice. We constructed a rabies-G pseudotyped lentiviral vector encoding the human VEGF gene (EIAV–VEGF) that produced VEGF after *in vitro* transduction of the D17 cell line (Fig. 1g). VEGF production was also confirmed by the detection of a clear green fluorescent protein (GFP) signal in spinal cord motor neurons two months after injection of EIAV–VEGF–IRES–GFP (a bicistronic EIAV vector genome which encodes human VEGF and GFP linked by the internal ribosome entry site (IRES) in a single transcription unit) into hindlimb muscles (Fig. 1h).

To explore the efficacy of VEGF gene transfer on motor neuron survival in the SOD1^{G93A} mice, EIAV–VEGF or EIAV–LacZ vectors were injected bilaterally into the hindlimb gastrocnemius, diaphragm, intercostal, facial and tongue muscles of SOD1^{G93A} mice before disease onset at three weeks of age. In SOD1^{G93A} mice injected with EIAV–LacZ, motor impairment—assessed by the rotarod task—was first detectable at 93 days of age but became rapidly more severe and very pronounced by 100 days of age (Fig. 2c, d). In contrast, the mice treated with EIAV–VEGF performed normally at 100 days of age and started to display significant motor performance deficits 28 days later (Fig. 2c, d). The delay in the onset of disease was highly significant in mice treated with EIAV–VEGF ($n = 7$) compared with the control animals injected with EIAV–LacZ ($n = 6$) (123 ± 11 versus 95 ± 4 ; $P < 0.0001$). Footprint analysis showed significant impairment in walking patterns of EIAV–LacZ-injected SOD1^{G93A} mice at 110 days and in the few mice that survived until 130 days of age (Fig. 2a, b). However, no obvious deficits were observed in EIAV–VEGF-treated mice, with stride length not being significantly different from wild-type mice at up to 130 days of age (Fig. 2a, b). The mice were killed at the clinical end point, defined as the time when they became unable to right themselves within 30 s, having been placed on their sides. Survival analysis in SOD1^{G93A} transgenic mice injected at three weeks of age with EIAV–LacZ or EIAV–VEGF showed that VEGF treatment significantly extended the lifespan by an average of 38 days—or increased life expectancy by as much as 30%—compared with EIAV–LacZ-injected animals (163 ± 15 days versus 125 ± 6 days; $P < 0.0001$; Fig. 2e). EIAV–VEGF gene therapy in SOD1^{G93A} mice therefore delayed the onset and slowed the progression of disease, because the duration of the disease was prolonged in EIAV–VEGF-treated mice by 40 days versus 30 days in EIAV–LacZ-injected animals. Furthermore, the level of VEGF protein within the spinal cord tissue at the end stage of disease was found to be significantly higher in EIAV–VEGF-treated mice (0.8 ± 0.05 ng per mg of tissue) than in EIAV–LacZ-treated control mice.

The therapeutic potential of VEGF was subsequently tested at the time of disease onset; this is a medical challenge, when considering that approximately half of the motor neurons have already died when clinical symptoms first appear. However, as treatment of ALS patients is at present initiated only after diagnosis of the disease, such an experiment is important. Ninety-day-old SOD1^{G93A} mice received bilateral injections of EIAV–VEGF or EIAV–LacZ into leg,

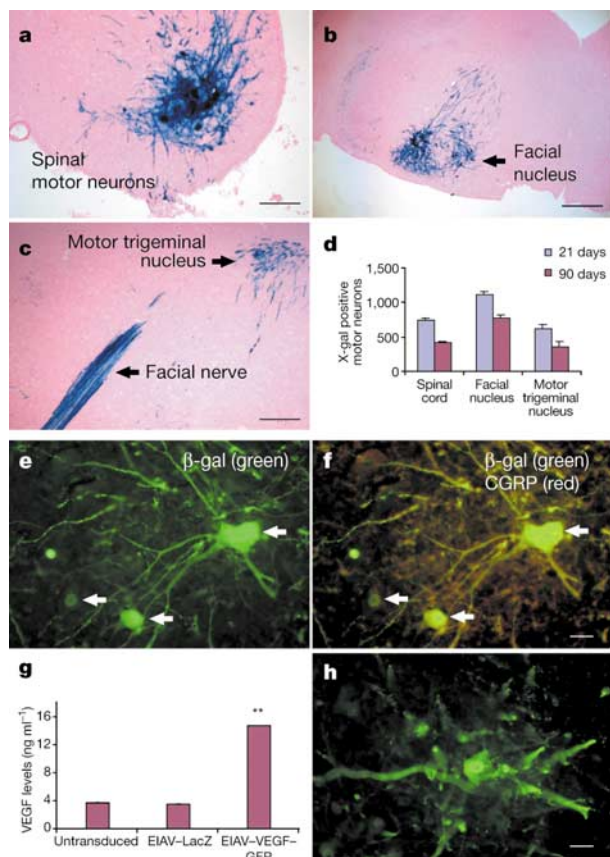


Figure 1 EIAV vector-mediated gene transfer *in vitro* and *in vivo*. **a–c**, Retrograde transport of the viral particles and transduction of the lumbar spinal cord and the brainstem motor neurons after injection of rabies-G pseudotyped EIAV–LacZ vector in the gastrocnemius (**a**) and the facial muscles (**b**, **c**). **d**, Cell counts of total transduced motor neurons in the spinal cord, facial nucleus and trigeminal motor nucleus in SOD1^{G93A} mice. **e**, **f**, Expression of β-gal (**e**) colocalizes with CGRP in spinal motor neurons to produce yellow staining (**f**). **g**, VEGF production in D17 cell seven days post-transduction. **h**, VEGF expression in spinal motor neurons after injection of EIAV–VEGF–GFP into hindlimb muscles in SOD1^{G93A} mice.

face, tongue, diaphragm and intercostal muscles. EIAV–LacZ-treated $SOD1^{G93A}$ mice survived to 127 ± 6 (mean $\pm$ s.d.) days of age ($n = 7$) (Fig. 2f). In contrast, $SOD1^{G93A}$ mice treated with EIAV–VEGF showed a statistically significant increase in lifespan to 146 ± 10 days of age ($n = 7$; $P < 0.0001$) (Fig. 2f). Neuromuscular function of treated mice was assessed using rotarod and footprint tests (Fig. 2g, h). Between 90 and 120 days of age, EIAV–LacZ-injected animals showed considerable decrease in motor performance (Fig. 2g). In contrast, VEGF treatment significantly slowed down the loss of motor performance (Fig. 2g). Footprint analysis showed that EIAV–LacZ-treated mice had motor weakness and dragged their legs, but these symptoms were substantially reduced by VEGF gene expression (Fig. 2h). Taken together, the above results indicate that VEGF treatment, even after onset of disease, induces not only an extended lifespan but also delays the functional deficits in this model. When injecting a similar amount of a rabies-G pseudotyped EIAV vector expressing glial-cell-line-derived neuro-

trophic factor (GDNF)—which has been previously reported to delay motor neuron loss in ALS mice^{17,18}—in 90-day-old $SOD1^{G93A}$ mice, we observed only a negligible increase in survival of six days (133 ± 4 ; $n = 6$; $P < 0.001$). EIAV–LacZ-treated $SOD1^{G93A}$ mice survived to 127 ± 6 days of age ($n = 7$). These findings indicate a superior specificity and potency of VEGF over GDNF for ALS therapy.

To correlate clinical evolution with histological data, the effect of VEGF treatment on the number of motor neurons in the spinal cord and facial nucleus was evaluated. The facial nucleus is known to be involved in the late stages of the disease in $SOD1^{G93A}$ mice¹⁶. The number of facial and spinal motor neurons was therefore compared in EIAV–LacZ and EIAV–VEGF-treated $SOD1^{G93A}$ mice at 115 days of age, as well as at the end stage of disease (Fig. 3). At the end stage of disease, we counted only $1,860 \pm 252$ facial motor neurons in EIAV–LacZ-treated $SOD1^{G93A}$ mice, but still as many as $2,680 \pm 157$ in EIAV–VEGF-treated animals ($n = 5$; $P < 0.001$), or a 44% increase in motor neuron survival (Fig. 3f). Histological evaluation of the lumbar spinal cord showed that retrograde delivery of VEGF in $SOD1^{G93A}$ mice induced a significant increase—as high as 125%—of spinal motor neurons as well (Fig. 3). Cell counts at 115 days of age showed that CGRP-positive motor neuron numbers per lumbar spinal cord section were significantly higher in EIAV–VEGF-treated than in EIAV–LacZ control mice ($n = 4$; $P < 0.001$; Fig. 3e). The lumbar motor neurons are normally affected early on in the course of the disease in $SOD1^{G93A}$ mice. Thus, as expected, at the preterminal end stage of disease, only minimal—though still statistically significant—differences in the lumbar motor neuron cell counts were detected between EIAV–LacZ- and EIAV–VEGF-treated animals (Fig. 3e).

Having established the therapeutic efficiency of EIAV–VEGF gene

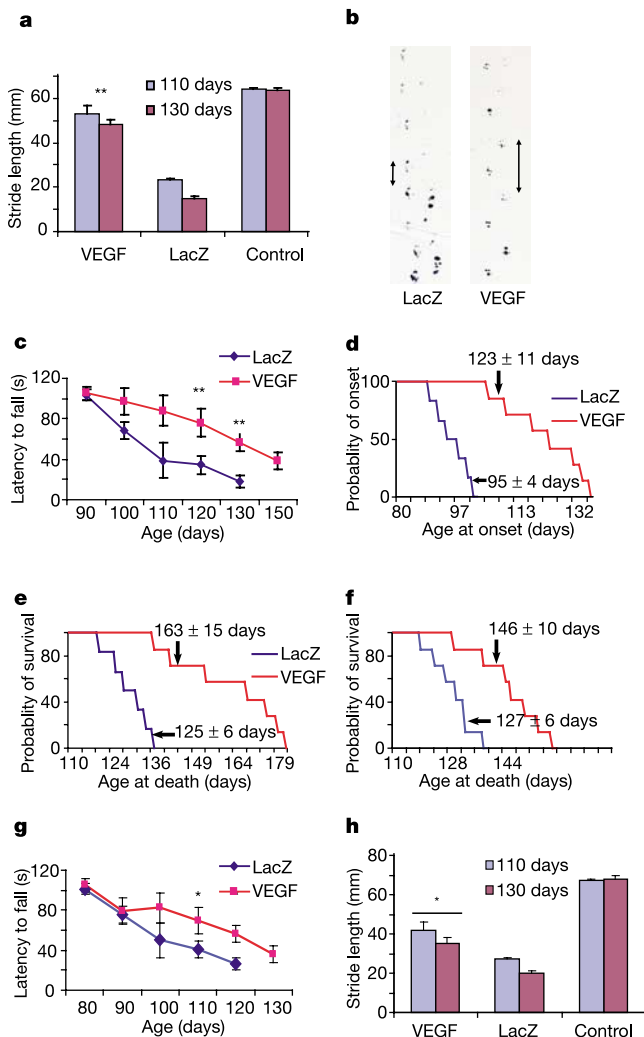


Figure 2 Effects of EIAV vector delivery of VEGF. Delivery of VEGF before (a–e) and at (f–h) onset of disease extends survival and delays motor deficits in $SOD1^{G93A}$ mice.

a, b, h. Footprints of 110- and 130-day-old EIAV–VEGF, EIAV–LacZ and wild-type mice. Footprint analysis showed impaired walking patterns in LacZ-injected mice compared with VEGF-injected and control animals. **b.** Representative footprints from EIAV–LacZ- and EIAV–VEGF-treated animals. **c, g.** Rotarod task of animals treated with EIAV–VEGF or EIAV–LacZ. **d.** Disease onset in $SOD1^{G93A}$ mice injected with EIAV–LacZ or EIAV–VEGF. **e, f.** Survival analysis in $SOD1^{G93A}$ mice injected at three weeks of age (e) and at 90 days of age (f) with EIAV vectors. * $P < 0.0001$, ANOVA (analysis of variance).

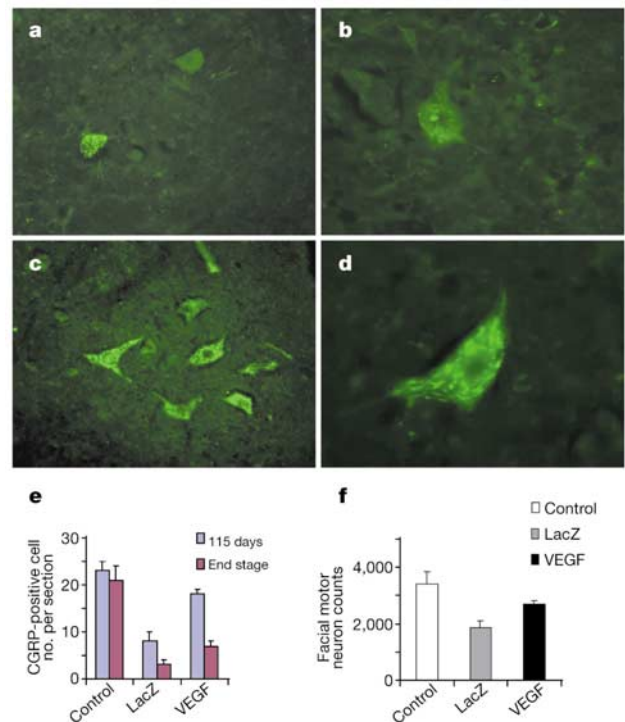


Figure 3 VEGF gene therapy protects spinal and brainstem motor neurons in $SOD1^{G93A}$ transgenic mice. **a–d.** Immunohistochemistry showing CGRP-positive motor neurons in lumbar spinal cord of EIAV–LacZ (a, b) and EIAV–VEGF injected animals (c, d). **e.** Cell counts of surviving lumbar spinal cord motor neurons in control (wild-type), EIAV–LacZ and EIAV–VEGF-treated $SOD1^{G93A}$ mice at 115 days of age (blue) and at the end stage of disease (red). **f.** Quantification of facial nucleus motor neurons in animals injected with EIAV–LacZ, EIAV–VEGF and control animals at the end stage of disease.

transfer, we assessed the effects of VEGF gene transfer on the architecture and histology of the spinal cord and brain. Upon macroscopic inspection of the spinal cord and brain after dissection of EIAV-VEGF-treated mice, we could not detect any abnormal signs. Microscopic quantification of glut-1 immunostained blood vessels in the spinal cord showed no significant changes in structure, size or density (Fig. 4c, d, g, h). In addition, immunostaining for the plasma protein albumin showed no detectable extravascular leakage of albumin after VEGF gene transfer, both in the spinal cord (Fig. 4a) and brain stem (Fig. 4b), compared with control animals (Fig. 4e, f). These results suggest that EIAV-VEGF gene transfer in SOD1^{G93A} mice provided significant protection against motor neuron death, without causing any prominent changes in the vascularization of the tissue. In addition, EIAV-mediated gene transfer in three-week-old SOD1^{G93A} mice induced minimal immune responses in muscle, as assessed using specific antibody markers to detect immune-responsive cells at the site of injection, five weeks after vector delivery (see Supplementary Information).

In summary, we have shown that retrogradely transported lentiviral vectors can be successfully used to transfer therapeutic genes to motor neurons for treatment of motor neuron degeneration. We have shown that VEGF is a potent neuroprotective factor with clear therapeutic effects in a model for motor neuron disease.

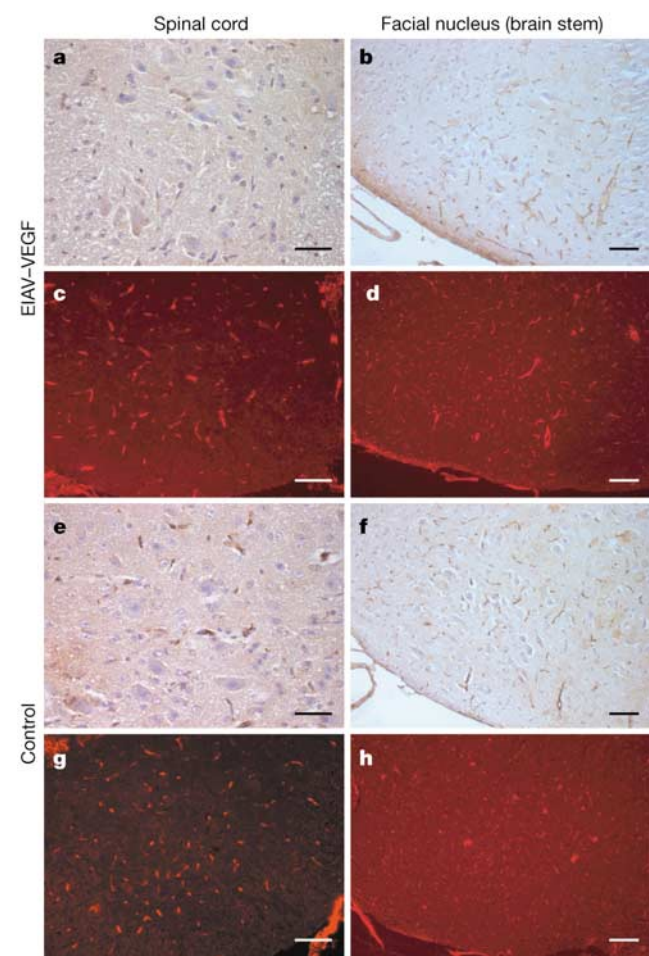


Figure 4 The effects of VEGF gene therapy five weeks after intramuscular injections of viral vectors. The effects of VEGF on vascularization in the spinal cord (**a, c, e, g**) (scale bar, 25 μ m) and in the brain stem (**b, d, f, h**) (scale bar, 100 μ m). **a, b, e, f**, Albumin staining in EIAV-VEGF-GFP (**a, b**) and control animals showing no vascular leakage after VEGF gene transfer (**e, f**). **c, d, g, h**, Glut-1 immunofluorescence in EIAV-VEGF-GFP-injected SOD1^{G93A} mice (**c, d**) and control group (**g, h**), showing no altered vascular volume and no abnormal vascular density after VEGF gene therapy.

Our data indicate that lentiviral-vector-mediated VEGF gene therapy achieves one of the greatest therapeutic effects on survival in the ALS mouse model reported until now. The effect was comparable to that reported for insulin-like growth factor (IGF-1) when delivered using an adeno-associated viral vector in the same mouse model^{17,18}. However, in contrast to this adeno-associated viral study, which also showed good efficacy with GDNF, we found GDNF to have very little effect on motor neuron survival and clinical disease progression. Given that IGF-1 activates expression of the VEGF gene through its effect on the hypoxia inducible transcription factor, HIF-1 (ref. 19), it is possible that the therapeutic molecule is in fact VEGF itself, which would explain the apparent similarity in therapeutic effects between VEGF and IGF-1 (ref. 20). The mechanism of the action of VEGF in this model remains to be characterized, but the data presented here suggest that it may operate by means of a direct effect on motor neurons rather than an indirect consequence of alterations in the vasculature of the spinal cord and brain.

Perhaps the most important observation in our study was that VEGF retained its protective efficacy even when EIAV-VEGF was delivered after half of the motor neurons had died. The therapeutic efficacy of late VEGF delivery is relevant to clinical application in human disease, because ALS occurs sporadically without prior family history and thus cannot be diagnosed before onset in >90% of cases. Finally, our data offer the prospect of an improved treatment strategy for motor neuron diseases. Typically, delivery of recombinant growth factors such as GDNF and IGF-1 has resulted in marginal therapeutic efficacy in clinical trials, presumably because of their short half-life and poor access to motor neurons²¹. The present methodology circumvents such limitations. In addition to potentially solving the delivery problem, it is also likely that the sustained localized expression, made possible by EIAV-mediated gene therapy, may be critical for these neurotrophic factors to achieve their therapeutic potential. Given that EIAV vectors have been known to give stable, long-term expression of therapeutic molecules for as long as they have been examined (12 months so far)^{11,22,23}, there could be no need for repeat injections, making this an effective method for managing a chronic disease state.

Our data show, using behavioural and anatomical outcomes, that an EIAV lentiviral vector expressing VEGF induces a substantial increase in survival in a clinically relevant model of ALS. We believe that this approach may have potential as a safe and practical treatment for many of the motor symptoms of human ALS. □

Methods

Viral production

EIAV self-inactivating vector genomes were constructed from pONY8.0Z or pONY8.0G vectors described previously^{11,22}. The complementary DNA coding for the reporter gene *LacZ* or the human VEGF165 was cloned in the EIAV transfer vector and EIAV-VEGF-IRES-GFP has also been generated. Viral vector stocks pseudotyped with rabies-G glycoprotein were prepared using the HEK293T transient system as previously described^{11,24}. The titres ($\sim 1 \times 10^9$ TU ml⁻¹) of concentrated EIAV-LacZ viral vectors were estimated by transduction of D17 cells. The titres ($\sim 7 \times 10^8 - 3 \times 10^9$ TU ml⁻¹) of the EIAV-VEGF or EIAV-VEGF-IRES-GFP vectors were estimated using real-time quantitative polymerase chain reaction with reverse transcription (RT-PCR) by comparison to EIAV-LacZ vectors and normalized for viral RNA^{25,26}.

Animal model

Transgenic mice overexpressing human SOD1 carrying a Gly93-Ala mutation were used⁷. This line of mice has the high-expressing form of mutant SOD1 and animals develop disease onset at about 90 days of age and die about 30 days later. Transgenic progeny were identified by PCR using primers specific for human SOD1 (ref. 7).

Viral vector delivery

Rabies-G pseudotyped lentiviral vectors carrying human VEGF or *LacZ* genes were injected bilaterally into the hindlimb gastrocnemius, facial, diaphragm, tongue and intercostal muscles of SOD1 transgenic mice before and at disease onset. The first group of SOD1 mice received injections of EIAV-VEGF-IRES-GFP ($n = 7$) at 21 days of age. The control group was treated with EIAV-LacZ vector ($n = 6$). The second set of animals was injected at the onset of disease (90-day-old mice) with EIAV-VEGF ($n = 7$) and the

EIAV–LacZ control ($n = 7$). Each mouse was injected with a total dosage of 90 μ l of viral solution. Six sites per hindlimb muscle were injected with 5 μ l per site.

Behavioural analysis

We analysed a rotarod task of the SOD1 mice by an Economex Rotarod instrument (Colombus Instruments) every ten days during the light phase of the 12 h light/12 h dark cycle. We performed three trials, and recorded the longest duration on the rod for every mouse. The timer is stopped when the mice fall from the rod or after an arbitrary limit of 180 s. Footprint analysis was also performed. Mouse hind paws were covered with ink to record walking patterns during continuous locomotion, and stride length was measured.

Histology and immunohistochemistry

Animals were perfused transcardially with 0.9% NaCl solution followed by ice-cold 4% paraformaldehyde. Spinal cord, brain and muscle tissues were dissected out and post-fixed overnight in the same solution and then transferred to 30% sucrose. Tissues were analysed by immunohistochemistry and X-gal (5-bromo-4-chloro-3-indolyl- β -D-galactoside) reaction. Spinal cord and brain were cut on a sliding cryostat microtome (Leica) at a thickness of 20 μ m and collected as free-floating sections in PBS containing sodium azide. Primary antibodies were used as follows: rabbit anti- β -gal (1:1,000, Europe BioProducts); rabbit anti-CGRP (1:3,000, Sigma); mouse anti-Chat (1:1,000, gift from B. K. Hartman and C. Cozari); CD3-T cells (1:50, Dako); P7/7-MHCII (1:100, Dako); goat anti-Glut-1 (1:20, Santa Cruz Biotechnology); and rabbit anti-albumin (1:250, ICN/Cappel).

VEGF ELISA

Dog osteosarcoma D17 cells were transduced in the presence of 8 mg ml⁻¹ polybrene as described previously²⁴. Cells were transduced with either EIAV–VEGF or EIAV–LacZ vectors. Transduced cells were passaged three times before analysis of transgene expression. One week post-transduction supernatants were collected and the VEGF levels measured by enzyme-linked immunosorbent assay (ELISA) (R&D Systems). The protein amino-acid sequences of dog VEGF are highly homologous to the human VEGF protein used in this kit, which has been used previously to determine dog VEGF levels²⁷. To determine plasma VEGF levels, we collected blood in 10-ml vacuum tubes containing 100 μ l of a 4% tri-sodium citrate solution, quickly centrifuged it and stored plasma fractions at -80 °C until analysis. VEGF ELISA assay measurements were also carried out using tissue samples from spinal cord and brain stem.

Received 11 February; accepted 5 April 2004; doi:10.1038/nature02544.

- Mulder, D. W. Clinical limits of amyotrophic lateral sclerosis. *Adv. Neurol.* **36**, 15–22 (1982).
- Carmeliet, P. Blood vessels and nerves: common signals, pathways and diseases. *Nature Rev. Genet.* **4**, 710–720 (2003).
- Schratzberger, P. *et al.* Favorable effect of VEGF gene transfer on ischemic peripheral neuropathy. *Nature Med.* **6**, 405–413 (2000).
- Jin, K. L., Mao, X. O. & Greenberg, D. A. Vascular endothelial growth factor: direct neuroprotective effect in *in vitro* ischemia. *Proc. Natl Acad. Sci. USA* **97**, 10242–10247 (2000).
- Oosthuyse, B. *et al.* Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nature Genet.* **28**, 131–138 (2001).
- Lambrechts, D. *et al.* VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nature Genet.* **34**, 383–394 (2003).
- Gurney, M. E. *et al.* Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation. *Science* **264**, 1772–1775 (1994).
- Brujin, L. I. *et al.* ALS-linked SOD1 mutant G85R mediates damage to astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions. *Neuron* **18**, 327–338 (1997).
- Azzouz, M. *et al.* Increased motoneuron survival and improved neuromuscular function in transgenic ALS mice after intraspinal injection of an adeno-associated virus encoding Bcl-2. *Hum. Mol. Genet.* **9**, 803–811 (2000).
- Cleveland, D. W. & Rothstein, J. D. From Charcot to Lou Gehrig: deciphering selective motor neuron death in ALS. *Nature Rev. Neurosci.* **2**, 806–819 (2001).
- Mazarakis, N. D. *et al.* Rabies virus glycoprotein pseudotyping of lentiviral vectors enables retrograde axonal transport and access to the nervous system after peripheral delivery. *Hum. Mol. Genet.* **10**, 2109–2121 (2001).
- Wong, L. F. *et al.* Transduction patterns of pseudotyped lentiviral vectors in the nervous system. *Mol. Ther.* **9**, 101–111 (2004).
- Rosen, D. R. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **364**, 362 (1993).
- Ferreirinha, F. *et al.* Axonal degeneration in paraplegin-deficient mice is associated with abnormal mitochondria and impairment of axonal transport. *J. Clin. Invest.* **113**, 231–242 (2004).
- Sagot, Y., Rosse, T., Vejsada, R., Perrelet, D. & Kato, A. C. Differential effects of neurotrophic factors on motoneuron retrograde labeling in a murine model of motoneuron disease. *J. Neurosci.* **18**, 1132–1141 (1998).
- Dal Canto, M. C. & Gurney, M. E. Neuropathological changes in two lines of mice carrying a transgene for mutant human Cu,Zn SOD, and in mice overexpressing wild type human SOD: a model of familial amyotrophic lateral sclerosis (FALS). *Brain Res.* **676**, 25–40 (1995).
- Wang, L. J. *et al.* Neuroprotective effects of glial cell line-derived neurotrophic factor mediated by an adeno-associated virus vector in a transgenic animal model of amyotrophic lateral sclerosis. *J. Neurosci.* **22**, 6920–6928 (2002).
- Kaspar, B., Llado, J., Sherkat, N., Rothstein, J. & Gage, F. Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* **301**, 839–842 (2003).
- Fukuda, R. *et al.* Insulin-like growth factor 1 induces hypoxia-inducible factor 1-mediated vascular endothelial growth factor expression, which is dependent on MAP kinase and phosphatidylinositol 3-kinase signaling in colon cancer cells. *J. Biol. Chem.* **277**, 38205–38211 (2002).
- Hellstrom, A. *et al.* Low IGF-1 suppresses VEGF-survival signaling in retinal endothelial cells: direct correlation with clinical retinopathy of prematurity. *Proc. Natl Acad. Sci. USA* **98**, 5804–5808 (2001).
- Mitchell, J. D., Wokke, J. H. & Borasio, G. D. Recombinant human insulin-like growth factor I

- (rhIGF-I) for amyotrophic lateral sclerosis/motor neuron disease. *Cochrane Database Syst. Rev.* **3**, CD002064 (2002).
- Azzouz, M. *et al.* Multicentric lentiviral vector-mediated striatal gene transfer of aromatic L-amino acid decarboxylase, tyrosine hydroxylase, and GTP cyclohydrolase I induces sustained transgene expression, dopamine production, and functional improvement in a rat model of Parkinson's disease. *J. Neurosci.* **22**, 10302–10312 (2002).
- Bienemann, A. S. *et al.* Long-term replacement of a mutated nonfunctional CNS gene: reversal of hypothalamic diabetes insipidus using an EIAV-based lentiviral vector expressing arginine vasopressin. *Mol. Ther.* **7**, 588–596 (2003).
- Mitrophanous, K. *et al.* Stable gene transfer to the nervous system using a non-primate lentiviral vector. *Gene Ther.* **6**, 1808–1818 (1999).
- Rohll, J. B. *et al.* Design, production, safety, evaluation, and clinical applications of nonprimate lentiviral vectors. *Methods Enzymol.* **346**, 466–500 (2002).
- Martin-Rendon, E., White, L. J., Olsen, A., Mitrophanous, K. A. & Mazarakis, N. D. New methods to titrate EIAV-based lentiviral vectors. *Mol. Ther.* **5**, 566–570 (2002).
- Mohammed, S. I. *et al.* Effects of the cyclooxygenase inhibitor, piroxicam, on tumor response, apoptosis, and angiogenesis in a canine model of human invasive urinary bladder cancer. *Cancer Res.* **62**, 356–358 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Mehmet and A. Kingsman for critical review of the manuscript. SOD1G93A transgenic mice were a gift from P. Shaw. Oxford BioMedica supported this work.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to M.A. (m.azzouz@oxfordbiomedica.co.uk) or N.D.M. (n.mazarakis@oxfordbiomedica.co.uk).

Premature ageing in mice expressing defective mitochondrial DNA polymerase

Aleksandra Trifunovic^{1,2}, Anna Wredenberg^{1,2}, Maria Falkenberg¹, Johannes N. Spelbrink³, Anja T. Rovio³, Carl E. Bruder⁴, Mohammad Bohlooly-Y⁴, Sebastian Gidlöf^{1,2}, Anders Oldfors⁵, Rolf Wibom⁶, Jan Törnelli⁴, Howard T. Jacobs³ & Nils-Göran Larsson^{1,2}

¹Department of Medical Nutrition and ²Department of Biosciences, Karolinska Institutet, Novum, Karolinska University Hospital, S-141 86 Stockholm, Sweden

³Institute of Medical Technology and Tampere University Hospital, FIN-33014 University of Tampere, Finland

⁴Astra Zeneca R&D, S-431 83 Mölndal, Sweden

⁵Department of Pathology, Sahlgrenska University Hospital, S-413 45 Göteborg, Sweden

⁶Department of Laboratory Medicine, Karolinska Institutet, Karolinska University Hospital, S-141 86 Stockholm, Sweden

Point mutations and deletions of mitochondrial DNA (mtDNA) accumulate in a variety of tissues during ageing in humans¹, monkeys² and rodents³. These mutations are unevenly distributed and can accumulate clonally in certain cells, causing a mosaic pattern of respiratory chain deficiency in tissues such as heart⁴, skeletal muscle⁵ and brain⁶. In terms of the ageing process, their possible causative effects have been intensely debated because of their low abundance and purely correlative connection with ageing^{7,8}. We have now addressed this question experimentally by creating homozygous knock-in mice that express a proof-reading-deficient version of PolgA, the nucleus-encoded catalytic subunit of mtDNA polymerase. Here we show that the knock-in mice develop an mtDNA mutator phenotype with a threefold to fivefold increase in the levels of point mutations, as well as increased amounts of deleted mtDNA. This increase in somatic mtDNA mutations is associated with reduced lifespan and premature onset of ageing-related phenotypes such as weight loss, reduced subcutaneous fat, alopecia (hair loss), kyphosis

(curvature of the spine), osteoporosis, anaemia, reduced fertility and heart enlargement. Our results thus provide a causative link between mtDNA mutations and ageing phenotypes in mammals.

The mtDNA polymerases of yeast and animal cells have an inherent 3'-5' exonuclease activity necessary for proof-reading newly synthesized mtDNA⁹⁻¹¹. The exonuclease activity of the *Saccharomyces cerevisiae* mtDNA polymerase, Mip1p, is critically dependent on specific, conserved aspartate residues located in three 'exonuclease' domains^{10,11}. We created knock-in mice (Fig. 1a-d) expressing mtDNA polymerase with an altered catalytic PolgA subunit that contained an alanine instead of the critical aspartate residue of the second exonuclease domain (D257A). We first generated the *PolgA^{mutNeo}* allele (Fig. 1a) in embryonic stem cells (Fig. 1b) and obtained germline transmission of this allele (Fig. 1c, d). Next, we excised the *Frt*-site-flanked neomycin-resistance gene by mating heterozygous *PolgA^{mutNeo}* mice with a line ubiquitously expressing *Flp*-recombinase¹², thus generating heterozygous *PolgA^{mut}* (+/*PolgA^{mut}*) mice (Fig. 1a, c, d). We chose to perform this *in vivo* excision procedure because others have shown that the presence of the neomycin resistance gene in a targeted locus can result in a hypomorphic allele¹³. Finally, we performed intercrosses of +/*PolgA^{mut}* mice to generate homozygous knock-in mice (*PolgA^{mut}/PolgA^{mut}*), hereafter denoted 'mtDNA-mutator' mice.

We analysed recombinant mtDNA polymerase holoenzyme complexes *in vitro* and found that mtDNA polymerase containing a PolgA subunit bearing the Asp → Ala exonuclease domain mutation had a profound reduction of exonuclease activity (Fig. 1e). Others have shown that certain exonuclease-deficient versions of yeast mtDNA polymerase have reduced capacity to synthesize DNA^{10,11}. However, no decrease in DNA polymerase activity was observed in biochemical assays with recombinant exonuclease-deficient mtDNA polymerase (Fig. 1f) or in mitochondrial extracts from heterozygous knock-in (+/*PolgA^{mut}*) or mtDNA-mutator mice (Fig. 1g).

The mtDNA-mutator mice had a normal appearance until the age of ~25 weeks, when we first noted slight kyphosis and alopecia. As the animals got older the kyphosis became marked and varying degrees of alopecia appeared (Fig. 2a). The median lifespan of the mtDNA-mutator mice was ~48 weeks and all of them died before the age of 61 weeks (Fig. 2e). Weight gain of the mtDNA-mutator mice started to decelerate from ~15–20 weeks of age and weight loss was noted from ~24 weeks of age (Fig. 2f, g). Alopecia is common in human ageing¹⁴. In addition, a decrease in body weight occurs in humans older than 60 years¹⁵ and is also well documented in mice older than ~1.5 years¹⁶.

We performed quantitative assessments of body composition with X-ray densitometry of the whole mouse (Fig. 2c) or dissected

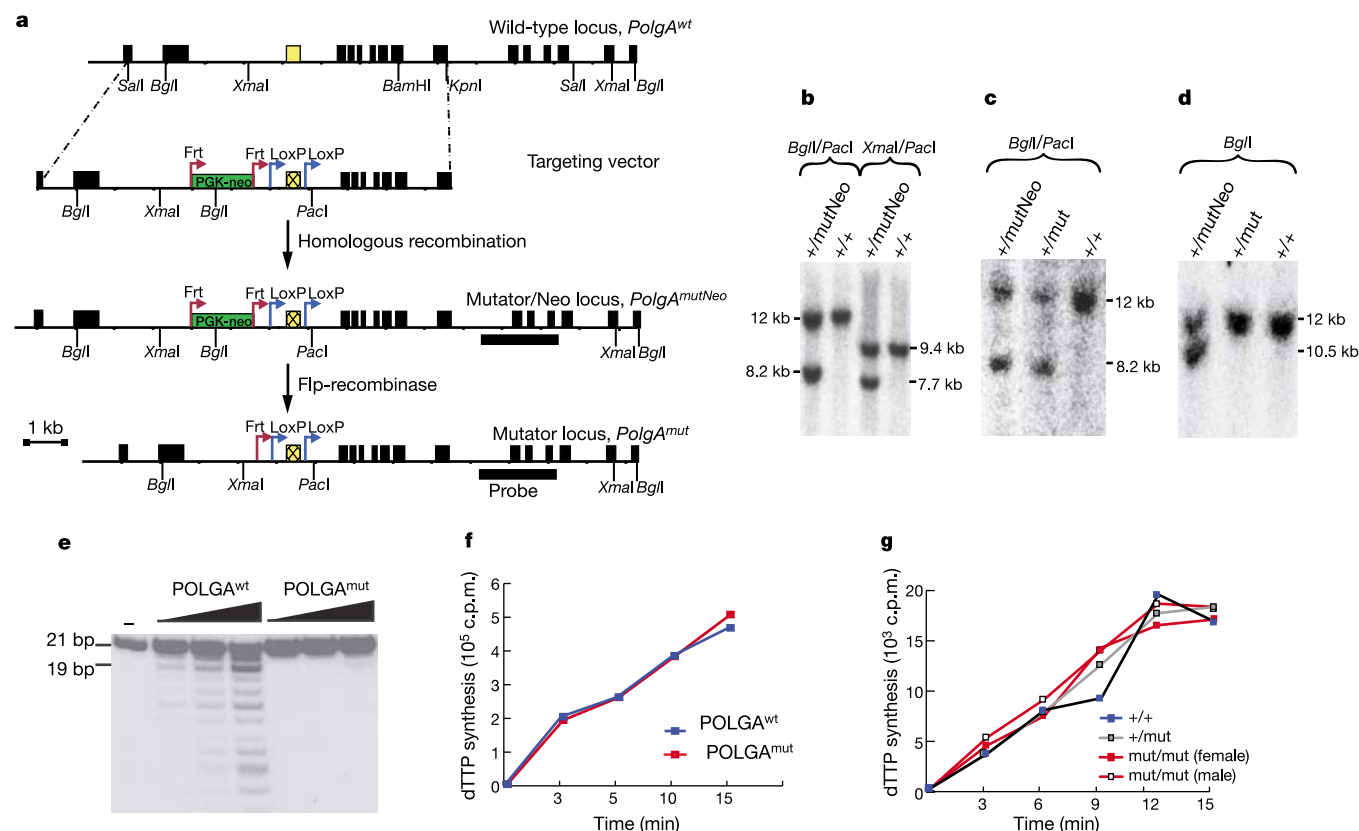


Figure 1 Creation of mtDNA-mutator mice and biochemical assays of DNA polymerase and exonuclease activities. **a**, Restriction map of the 5' region of the wild-type *PolgA* locus (*PolgA^{wt}*), the targeting vector, the *PolgA* locus after homologous recombination (*PolgA^{mutNeo}*) and the mutator locus obtained after *Flp*-recombinase-mediated excision of the *PGK-neo* gene (*PolgA^{mut}*). The sequence of exon 3 (yellow box) of the targeting vector was changed to encode an alanine instead of the highly conserved aspartate residue. The thick black line indicates the probe used to screen for homologous recombination. **b**, Digestion of ES cell DNA with the restriction endonucleases *Bgl*I/*Pac*I or *Xma*I/*Pac*I generates a novel fragment of ~8.2 kb and ~7.7 kb, respectively, in clones that are heterozygous for *PolgA^{mutNeo}*. **c, d**, Mice heterozygous for *PolgA^{mutNeo}* (+/*mutNeo*) were

mated to *Flp*-mice to excise the *Frt*-flanked *PGK-neo* cassette to obtain heterozygous *PolgA^{mut}* mice (+/*mut*). **e**, Biochemical assays of exonuclease activity. Recombinant human POLGA proteins were co-expressed with the human accessory subunit, POLGB, to obtain recombinant mtDNA polymerase holoenzyme. The enzyme containing the POLGA^{wt} polypeptide exhibits exonuclease activity, whereas the POLGA^{mut}-containing enzyme lacks this activity. **f**, DNA synthesis activities with recombinant holoenzymes consisting of human POLGA^{wt} or POLGA^{mut} and POLGB. **g**, DNA-synthesis activities in mitochondrial extracts isolated from liver of wild-type (+/+), heterozygous *PolgA^{mut}* mice (+/*mut*) and mtDNA-mutator mice (*PolgA^{mut}/PolgA^{mut}*, *mut/mut*).

femur bones (Fig. 2d). The ratio of lean body mass to length was normal, whereas the fat content was reduced in mtDNA-mutator mice (Fig. 2c), consistent with their overall appearance (Fig. 2a) and histological analyses of subcutaneous fat (Fig. 2b). In humans, body fat decreases with age after 65 years and a reduction in subcutaneous fat is common in the ageing skin¹⁴. Measurements of whole-body bone mineral density (BMD) showed a clear reduction at the age of 40 weeks in mtDNA-mutator mice (Fig. 2c), consistent with the clinical features of osteoporosis (marked kyphosis; Fig. 2a). There was a tendency to reduction in the ratio of whole-body bone mineral content (BMC) to length at 40 weeks of age (Fig. 2c). We

also analysed dissected femur, to improve the precision of the BMD and BMC measurements, and found no difference between mtDNA-mutator and wild-type mice at 20 weeks of age. However, there was a clear reduction in BMD (Fig. 2d), BMC (not shown) and BMC/length (Fig. 2d) in femur of mtDNA-mutator mice at the age of 40 weeks. In conclusion, the X-ray densitometry revealed a change in body composition, with reduced fat content and development of osteoporosis in the mtDNA-mutator mice. Human ageing is similarly accompanied by kyphosis and osteoporosis¹⁴. Osteoporosis is also found in ageing rodents as demonstrated by decreased bone densities in aged wild-type mice¹⁷.

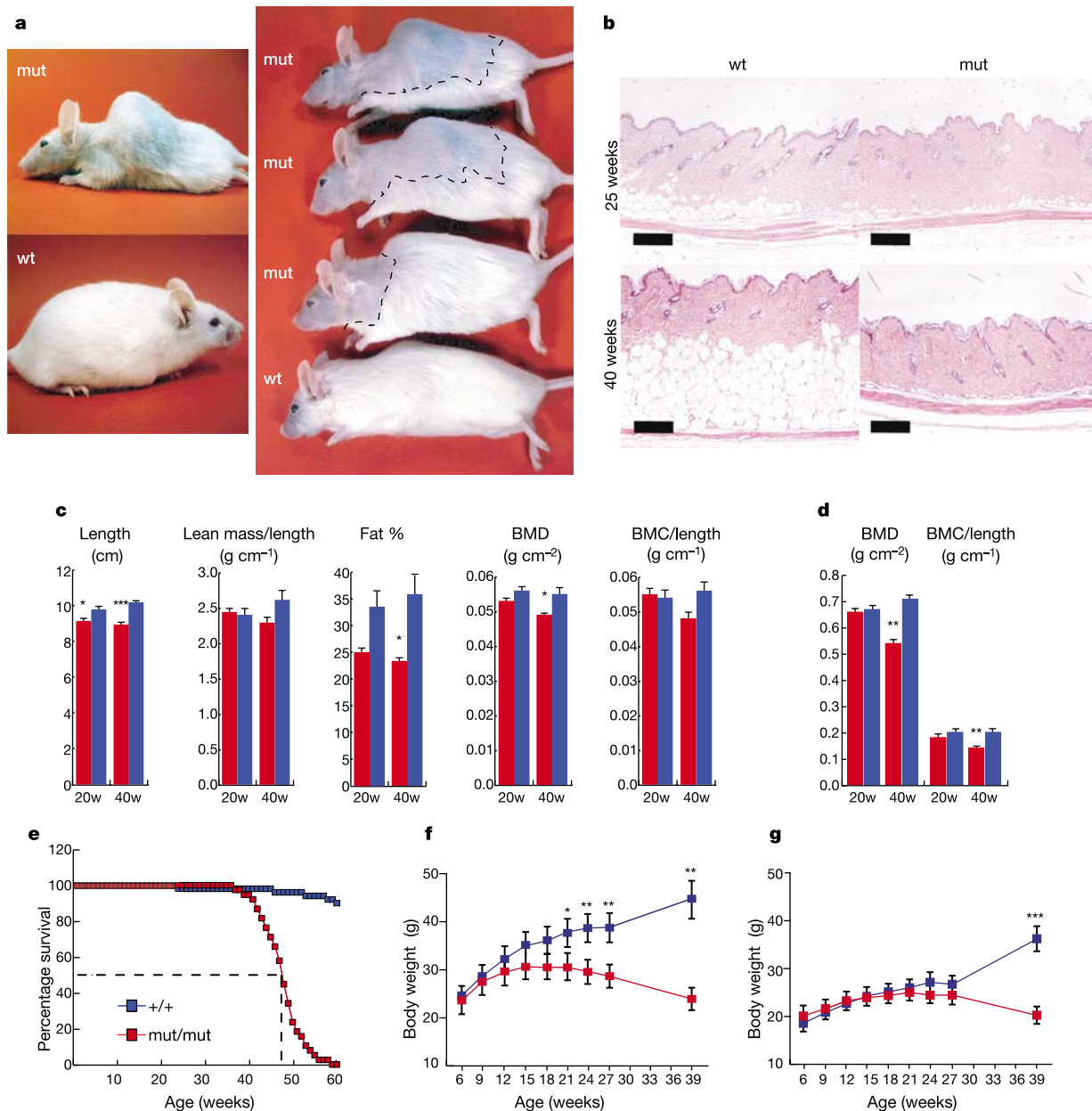


Figure 2 Ageing-related phenotypes observed in mtDNA-mutator mice. **a**, The mtDNA-mutator mice show reduced body size, kyphosis, reduced hair density and different stages of alopecia (delineated by dashed line) in comparison with littermate wild-type (wt) mice at the age of ~40–45 weeks. **b**, Haematoxylin- and eosin-stained cross-section of dorsal skin from 25- and 40-week-old wild-type (wt; $n = 3$ animals at each time point) and mtDNA-mutator (mut; $n = 3$) mice. Scale bar, 100 μ m. **c**, Quantitative assessment of body length, lean mass, fat content, BMD and BMC by X-ray densitometry of mtDNA-mutator mice (red bars) and wild-type littermates (blue bars). **d**, Quantitative

assessment of BMD and BMC by X-ray densitometry of dissected femur from mtDNA-mutator mice (red bars) and wild-type littermates (blue bars). 20w, 20 weeks; 40w, 40 weeks. **e**, Survival curves for wild-type ($n = 50$; blue line) and mtDNA-mutator ($n = 38$; red line) mice. **f**, Body weight curves for male wild-type ($n = 5$ –13 mice at the different time points; blue line) and mtDNA-mutator ($n = 8$ –10; red line) mice. **g**, Body weight curves for female wild-type ($n = 5$ –7; blue line) and mtDNA-mutator ($n = 9$ –11; red line) mice. Asterisk, $P < 0.05$; double asterisk, $P < 0.01$; triple asterisk, $P < 0.001$, Student's t -test. All error bars indicate s.e.m.

We found significantly lower haemoglobin concentrations in peripheral blood ($P < 0.001$, Student's *t*-test) of mtDNA-mutator mice (81 ± 3 g per litre, mean $\pm$ s.e.m; $n = 9$ analysed animals) in comparison with wild-type littermates (106 ± 3 g per litre; $n = 8$) at the age of 25 weeks. The anaemia in mtDNA-mutator mice was macrocytic and hypochromic (Supplementary Table 1). In addition, we observed extramedullary haematopoiesis in the liver (Fig. 3b) and spleen enlargement (Supplementary Tables 2 and 3), with increased relative contribution of red pulp (Fig. 3a) in mtDNA-mutator mice. Increased haematopoiesis in the spleen and foci of haematopoiesis in the liver are common findings in ageing mice¹⁶. It should also be noted that anaemia of unknown aetiology is a frequent clinical problem in elderly humans¹⁸.

The heart weight in relation to body weight was increased in mtDNA-mutator mice (Supplementary Table 3). The 40-week-old mtDNA-mutator mice also had an enlarged lumen of the left ventricle of the heart (Fig. 3c). Increases of heart weight and left ventricle hypertrophy are normally found in the ageing human heart^{14,19}. Cardiomyopathy is common with increasing age in wild-type mice¹⁶. We performed enzyme histochemical staining of heart muscle tissue sections from mtDNA-mutator mice and found a

mosaic pattern, with cytochrome *c* oxidase deficiency in some cardiomyocytes (Fig. 3d), similar to findings in ageing human hearts⁴. In accordance with this finding, light microscopy of thin sections of heart muscle showed that some cardiomyocytes had a vacuolated appearance because of accumulations of enlarged mitochondria (Fig. 3e). Electron microscopy studies of tissue sections from heart confirmed that there was an accumulation of abnormal mitochondria in a subset of the cardiomyocytes (Fig. 3f) in mtDNA-mutator mice.

A profound reduction in fertility of mtDNA-mutator mice of both sexes was found. In matings between 15 mtDNA-mutator females and 15 wild-type males, 14 of the females did become pregnant, each giving birth to one or two litters of normal size. However, the females did not become pregnant again after the age of 20 weeks, despite being continuously exposed to males for several months. We also set up matings over several months between 8 mtDNA-mutator males and 16 wild-type females, but obtained only a single, small litter. Testes of the mtDNA-mutator males were considerably smaller than wild-type testes from 12 weeks of age onwards (Fig. 3g; Supplementary Tables 2 and 3). Inspection of testes from mtDNA-mutator mice revealed reduced sperm content

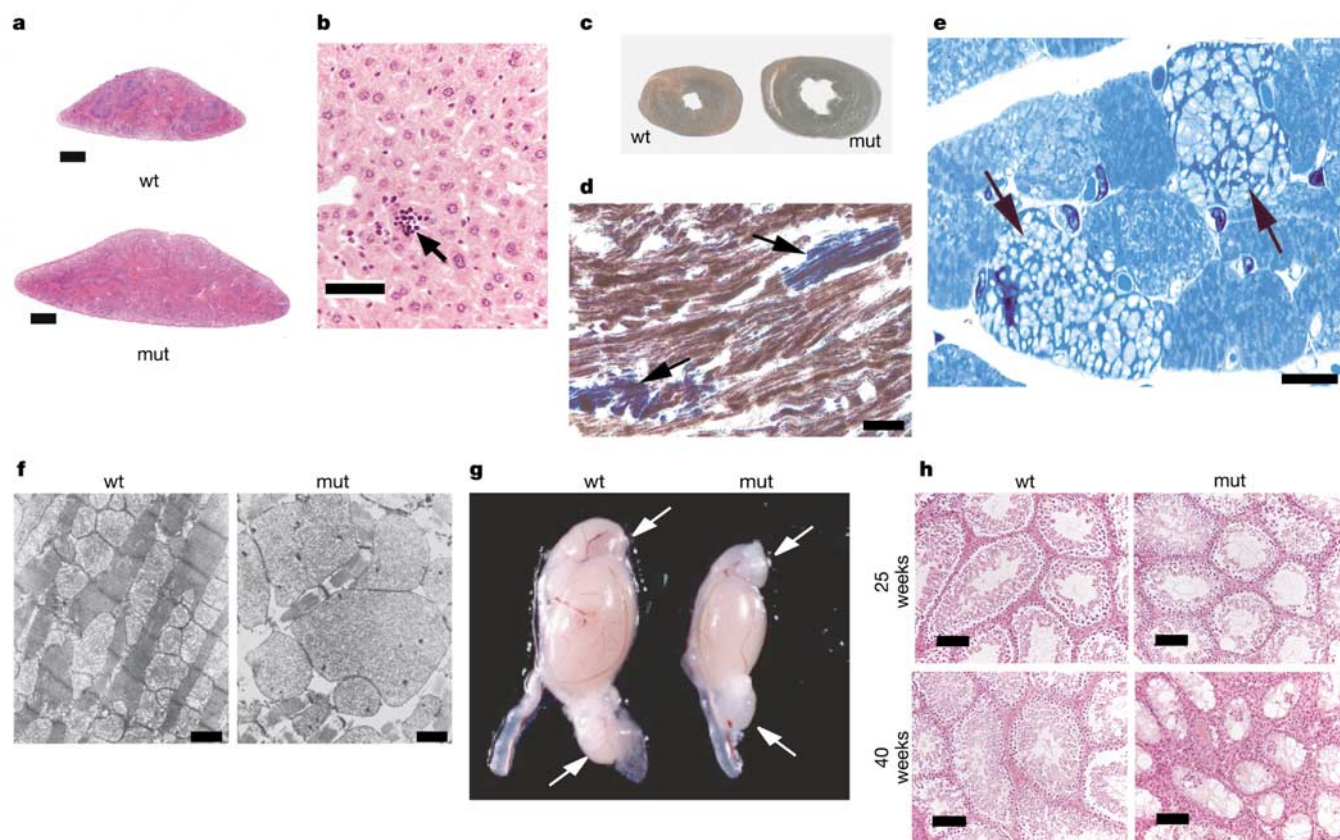


Figure 3 Histological analyses of different tissues from mtDNA-mutator mice.

a, Haematoxylin- and eosin-stained sections of spleen of wild-type (wt; $n = 8$ analysed animals) and mtDNA-mutator mice (mut; $n = 8$) at age 25 weeks. Scale bar, 0.5 mm. **b**, Haematoxylin- and eosin-stained section from the liver of mtDNA-mutator ($n = 8$) and wild-type ($n = 8$, not shown) mice at age 25 weeks. Extramedullary haematopoiesis (arrow) is present in mtDNA mutator mice. Scale bar, 50 μ m. **c**, Unstained thick sections of heart from wild-type ($n = 3$) and mtDNA-mutator mice ($n = 3$) at age 40 weeks. **d**, Enzyme histochemical double staining for cytochrome *c* oxidase (COX) and succinate dehydrogenase (SDH) activities in heart of an mtDNA-mutator mouse at age 40 weeks. The presence of both COX and SDH activity results in a brown colour. A mosaic pattern is present in the mtDNA-mutator mouse heart and some cardiomyocytes appear blue (arrows) because of a lack of COX activity (catalytic subunits are mtDNA encoded) and

preserved SDH activity (all subunits are nucleus encoded). Scale bar, 20 μ m. A total of three mtDNA-mutator and three wild-type mice were analysed. **e**, Light microscopic image of a thin section of Epon-embedded heart muscle from an mtDNA-mutator mouse at age 40 weeks shows cardiomyocytes with accumulation of enlarged mitochondria (arrows). Scale bar, 10 μ m. **f**, Electron micrographs of cardiomyocytes from wild-type ($n = 3$) and an mtDNA-mutator ($n = 3$) mice. Some cardiomyocytes of the mtDNA-mutator mouse contains enlarged abnormally shaped mitochondria. Scale bar, 1 μ m. **g**, Testis from wild-type ($n = 13$) and mtDNA-mutator ($n = 13$) mice at the age of 25 weeks. The size of the testis is reduced and no sperm is visible on macroscopic inspection of the epididymis (arrows) of the mtDNA-mutator mouse. **h**, Cross-sections of testis of wild-type ($n = 4-6$) and mtDNA-mutator ($n = 4-6$) mice at 25 and 40 weeks of age. Scale bar, 100 μ m.

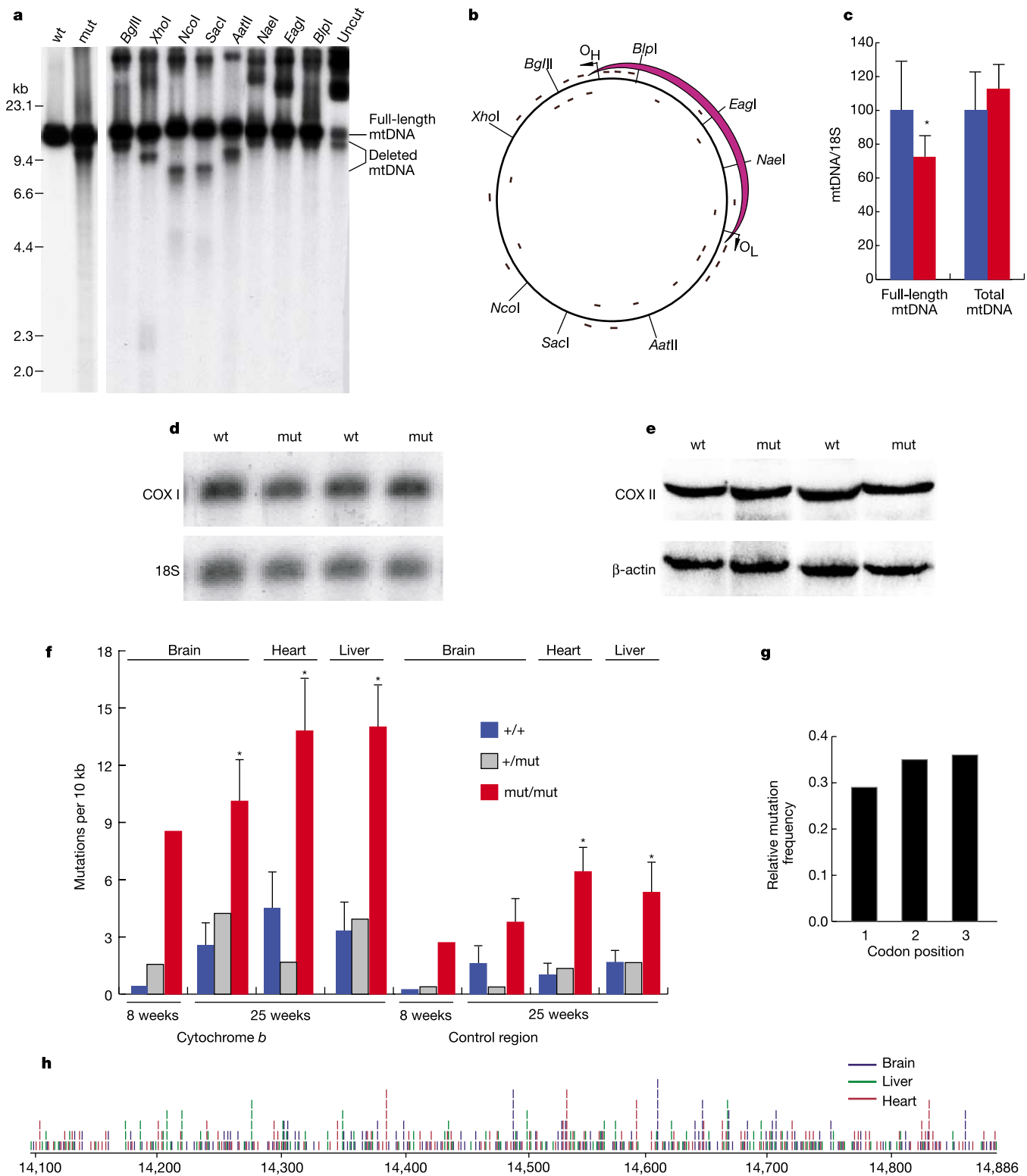


Figure 4 Analysis of mtDNA. **a**, Southern blot analysis of *Bgl*II-digested liver mtDNA from a wild-type and an mtDNA-mutator mouse at age 25 weeks (lane 1–2). Southern blot analysis of restriction-enzyme-digested mtDNA from an mtDNA-mutator mouse at age 25 weeks (lanes 3–11). **b**, The deleted mtDNA region is indicated with a purple arc. The black bars represent 29 strand-specific oligonucleotide probes used to map the mtDNA deletions. Oligonucleotides identical to the L-strand of mtDNA are inside and those identical to the H-strand are outside the circle. **c**, Quantification of mtDNA levels (mean $\pm$ s.e.m; asterisk, $P < 0.01$, Student's *t*-test) in liver from mtDNA-mutator (red bars; $n = 6$) and wild-type (blue bars; $n = 6$) mice at age 12–40 weeks. **d**, Northern blot analysis of cytochrome *c* oxidase subunit I (COX I) transcript levels in heart of wild-type and mtDNA-mutator mice at age 40 weeks. **e**, Western blot analysis of cytochrome *c*

oxidase subunit II polypeptide (COX II) levels in heart of wild-type and mtDNA-mutator mice at age 25 weeks. **f**, mtDNA mutation loads in mtDNA-mutator (mut/mut; red bars), heterozygous *PolgA*^{mut} (+/mut; grey bars) and wild-type (+/+; blue bars) mice. The bars indicate mean values. The standard deviation (indicated by error bars) was calculated in groups where DNA from three or four mice of the same genotype was sequenced. Asterisk, $P < 0.005$, Student's *t*-test. **g**, Relative mutation frequencies at all codon positions in the cytochrome *b* gene. **h**, Positions of individual mutations in the cytochrome *b* gene of brain (blue), heart (red) and liver (green) from 25-week-old mtDNA-mutator mice.

of the epididymis (Fig. 3g) and histological analyses of testis sections confirmed reduced sperm content in epididymis of mtDNA-mutator mice (data not shown). We observed severe testicular tubular degeneration with complete absence of sperm in 40-week-old mtDNA-mutator mice (Fig. 3h). In humans, female fertility decreases with age and males develop an age-associated decline in sperm count²⁰. Degenerated tubules may occupy as much as 60% of the testis in aged, wild-type mice²¹.

We performed Southern blot analyses and found a widespread tissue distribution of a class of shorter mtDNA molecules, up to ~12 kilobases (kb) in length, in mtDNA-mutator mice (Fig. 4a and Supplementary Fig. 1). Further analyses were performed with restriction enzymes that have single recognition sites in normal mouse mtDNA (Fig. 4a). The restriction enzymes *NaeI*, *EagI* and *BlnI* did not digest the shorter mtDNA molecules, whereas the restriction enzymes *BglII*, *XhoI*, *NcoI*, *SacI* and *AatII* did digest the shorter molecules and gave a DNA fragment pattern showing that the shorter species behaved as a linear, subgenomic—that is, deleted—mtDNA molecule (Fig. 4a). The amount of deleted mtDNA did not change over time and the levels of deleted mtDNA were comparable in all investigated tissues (Supplementary Fig. 1). Southern blot analyses with radiolabelled oligonucleotides or short DNA fragments (Supplementary Fig. 1) localized the commonly deleted region to the small arc between *O_H* and *O_L* (Fig. 4b), which is consistent with the results from the restriction enzyme analysis (Fig. 4a). The mean level of full-length mtDNA was ~70% of the level in wild-type mice (Fig. 4c). The reduced copy number of full-length mtDNA did not affect overall mtDNA expression; the steady-state levels of the mtDNA-encoded cytochrome *c* oxidase subunit 1 messenger RNA (Fig. 4d) and the mtDNA-encoded cytochrome *c* oxidase subunit 2 polypeptide (Fig. 4e) were normal. These findings are consistent with our previous observation that ~35–40% general reduction of mtDNA copy number in heterozygous mitochondrial transcription factor A knockout (+/*Tfam*⁻) mice has minimal effects on mtDNA expression and respiratory chain function²².

We cloned and sequenced segments of mtDNA from brain, heart and liver, and found that mtDNA-mutator mice contain an increased load (~3–5 times) of somatic mtDNA point mutations (Fig. 4f). The mutation background from PCR (polymerase chain reaction) was subtracted when the mtDNA mutation load was calculated. We also discarded the few multiple occurrences of a single mutation because they could reflect clonal expansion, which includes selection and/or genetic drift in addition to *de novo* mutagenesis. The presented frequency of somatic mtDNA mutations (Fig. 4f) thus represents a conservative minimal estimate. We combined the cytochrome *b* gene sequencing results from all tissues and time points of mtDNA-mutator mice (>600 individual mutations) and found that all codon positions were equally affected (Fig. 4g). In addition, the mutations were evenly distributed along the cytochrome *b* gene (Fig. 4h). There were thus no obvious mutational hot spots in mtDNA-mutator mice, although mutation load in the non-coding control region was lower than in the cytochrome *b* gene. The mutation load in heterozygous mice (Fig. 4f) was indistinguishable from that of wild-type littermates, and these mice displayed no phenotype, showing that the *PolgA*^{mut} mutation is recessive.

We further characterized respiratory chain function in the hearts of mtDNA-mutator mice (Supplementary Fig. 2) because of the findings of focal cytochrome *c* oxidase deficiency (Fig. 3d) and increased mitochondrial mass in some cardiomyocytes (Fig. 3e, f). We found slightly increased citrate synthase activities, consistent with a moderate increase in mitochondrial mass. There was a progressive reduction of respiratory chain enzyme activities and of mitochondrial ATP production rates (MAPR) in the hearts of mtDNA-mutator mice. The observed decline of respiratory chain enzyme activities and MAPR had profiles consistent with the

suggestion that the deficiencies were induced by mutations of mtDNA (Supplementary Fig. 2).

Our results establish a direct, experimental link between increased levels of somatic mtDNA mutations, respiratory chain dysfunction and phenotypes present in ageing mammals. The enhanced mtDNA mutation load of the mtDNA-mutator mouse is clearly generated somatically, because it is not shared with heterozygous or wild-type littermates. However, the detailed kinetics of accumulation of somatic mtDNA mutations remains to be elucidated. The mutation load is already substantial by 2 months, and is also rather uniform between tissues, suggesting that much of the mutation accumulation may occur during embryonic and/or fetal development. The onset of premature ageing is not accompanied, temporally, by a large *de novo* accumulation of mtDNA mutations around 6 months. Rather, it seems more plausible to ascribe it to cumulative physiological damage caused by the high mutation load during adult life, and/or to segregation or clonal expansion of specific mutations, as supported by the observed mosaicism for respiratory chain deficiency in the heart. Loss of vital cells in which mtDNA mutations have accumulated beyond a critical threshold (for example, through apoptosis or replicative senescence associated with telomere shortening brought on by oxidative stress) may be the critical process that manifests as premature ageing and reduced lifespan. We have previously generated a series of tissue-specific *Tfam* knockout mouse strains with reduced mtDNA expression and severe respiratory chain deficiency in different organs, accompanied by increased cell death and/or apoptosis^{23–26}. However, we found only minor or no induction of defence mechanisms against reactive oxygen species (ROS) in these knockouts^{24,27}. The possible involvement of increased ROS production as a mechanism of physiological damage leading to premature ageing in the mtDNA-mutator mouse should nevertheless not be discounted. Unlike tissue-specific *Tfam* knockouts, in which the effects on respiration and oxidative phosphorylation are essentially quantitative, the mtDNA-mutator creates an apparently random set of point mutations in genes for respiratory chain subunits. These may be expected to exhibit qualitative, as well as quantitative, abnormalities, of which increased ROS production and proton leak are likely consequences. The mtDNA-mutator mice will thus be a valuable tool for future experiments to determine whether the consequences of increased somatic mtDNA mutation can be counteracted by genetic, pharmacological or dietary interventions that affect ROS formation, bioenergetic homeostasis, cell death, apoptosis or other pathophysiological effects of respiratory chain dysfunction. Such approaches may allow us to design strategies to antagonize or delay deleterious consequences of naturally occurring somatic mtDNA mutations in human ageing. □

Methods

Creation of mtDNA-mutator mice

A genomic clone containing exons 1–16 of the mouse *PolgA* gene was isolated from a 129/SvJ λ Fix II phage library (Stratagene). A *Sall*–*KpnI* *PolgA* fragment of 8.1 kb (Fig. 1a) was cloned into pBluescript II SK+ (pBS, Stratagene) to generate plasmid pST1. A *Bam*HI–*Sall* fragment of 3.6 kb (Fig. 1a) was cloned into pBS to generate plasmid pST4. The inserts of plasmids pST1 and pST4 were completely sequenced. The plasmid pST8 was constructed by cloning a 1.4-kb *SacI* fragment of pST1 containing exon 3 (which encodes exonuclease domain 2) into pBS and used for site-directed mutagenesis with the oligonucleotides E3-1 (5'-gttagtggtggggcacaatgtttcctttgcccagccc-3') and E3-2 (5'-gggctcgggcaaggaacattgtgccccaccactaac-3') to change the aspartic acid codon (GAC) at position 257 to an alanine codon (GCC). The pDelboy plasmid (provided by D. Rossi) contains an *Frt*-site-flanked neomycin gene expressed from the phosphoglycerate kinase promoter (*PGK-neo*), two *loxP* sites and a cassette containing the herpes simplex virus thymidine kinase gene expressed from the phosphoglycerate kinase promoter (*PGK-HSV-TK*). The plasmid pST7 was constructed by exchanging one *loxP* site and the *PGK-HSV-TK* cassette in pDelboy with a DNA fragment containing a *SacI*–*loxP*–*PacI* sequence. The insert of pST8 was introduced into the *SacI*-site of pST7 to generate pST9. Plasmid pST6 was generated by replacing the 1.4-kb *SacI* fragment of pST1 with a DNA fragment containing a *KpnI*/EcoRI/*PacI* polylinker sequence. The final targeting vector for introducing the *PolgA*^{mutNeo} allele was obtained by ligating a 3.6-kb *XhoI*/*PacI* fragment from pST9 into pST6. The composition of DNA constructs was confirmed by restriction-

enzyme mapping and partial sequencing after each cloning step. The targeting vector was linearized with *SalI* and electroporated into 129 R1 embryonic stem (ES) cells. A total of 61 ES cell clones were subjected to Southern blot analysis and nine specifically targeted clones were found (Fig. 1b). We generated chimaeras by blastocyst injection of ES cells and obtained germline transmission of the *PolgA*^{mutNeo} allele from two clones. Mice heterozygous for the *PolgA*^{mutNeo} allele were mated with transgenic C57BL/6 mice ubiquitously expressing *Flp*-recombinase¹² to remove the *PGK-Neo* gene, thus generating heterozygous *PolgA*^{mut} mice (Fig. 1a, c, d). Heterozygous *PolgA*^{mut} mice (+/*PolgA*^{mut}) were intercrossed to generate mtDNA-mutator mice (*PolgA*^{mut}/*PolgA*^{mut}). This cross resulted in normal litter sizes (mean litter size 8.9 pups) and mendelian distributions of genotypes (*n* = 500 genotyped animals; wild type 25.8%, +/*PolgA*^{mut} 50.1% and *PolgA*^{mut}/*PolgA*^{mut} 24.1%) showing that homozygosity for *PolgA*^{mut} allele does not result in embryonic or early postnatal lethality. We followed the survival of wild-type (*n* = 50) and mtDNA-mutator (*n* = 38) mice (Fig. 2e). A total of 18 mtDNA-mutator mice died spontaneously, whereas the remaining 20 mice were killed in accordance with the ethical permit requirements because of moribund appearance. Standard Southern blot protocols were used for genotyping.

Analyses of DNA polymerase and exonuclease activities

A recombinant mtDNA polymerase holoenzyme was obtained by co-expression of human POLGA and POLGB subunits in the baculovirus system and biochemical purification. The DNA polymerase activities of recombinant proteins and mitochondrial extracts were determined by measuring incorporation of [α -³²P]dTTP into the artificial (rA).p(dT)₁₂₋₁₈ template. The exonuclease activities of recombinant proteins were determined by adding a substrate obtained by annealing a 5' ³²P-labelled 21-base oligonucleotide (5'-TGATGCTGCTGAGTCTGACTG-3') to M13mp18 single-stranded DNA (BioLabs). This substrate consists of a 20-base-pair double-stranded region with a one-nucleotide 3' mismatch. The exonuclease reactions were analysed by electrophoresis in 3 M urea/25% polyacrylamide gels.

Quantification of lean body mass, body fat, BMD and BMC

Mice were killed by CO₂ inhalation and densitometry was performed by using a PIXImus imager (GE Lunar). A total of three wild-type and three mtDNA-mutator mice were analysed at the age of 20 weeks, and five wild-type and five mtDNA-mutator mice were analysed at the age of 40 weeks. Body fat (percentage), lean body mass, BMD (g cm⁻²) and BMC (g) were determined in the whole mouse. BMD and BMC were also determined in the dissected femur bone. Field calibration and calibration versus the quality control phantom were performed each day before mouse or femur imaging.

Analysis of mtDNA mutations

Total DNA was extracted from the brains of 2-month-old wild-type, heterozygous and mtDNA-mutator mice. Enriched mtDNA was extracted from the isolated mitochondria (prepared by differential centrifugation) of different tissues of 6-month-old mice. The somatic mtDNA mutation load was determined by PCR, cloning, and sequencing, as described earlier²⁸, using primers that specifically amplified the cytochrome *b* gene (nucleotide pair 14,073–14,906) and non-coding control region (15,357–138) of mouse mtDNA²⁹ (GenBank NC_005089) together with flanking sequences.

Histology and biochemical analyses

Different tissues were fixed in buffered formalin, embedded in paraffin, sectioned and stained with haematoxylin and eosin. Enzyme histochemical and ultrastructural analyses of myocardium were performed as described^{25,26}. The measurement of respiratory chain enzyme complex activities, citrate synthase activity and MAPR was performed as previously described^{25,26,30}.

Received 16 February; accepted 29 March 2004; doi:10.1038/nature02517.

- Corral-Debrinski, M. *et al.* Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. *Nature Genet.* **2**, 324–329 (1992).
- Schwarze, S. R. *et al.* High levels of mitochondrial DNA deletions in skeletal muscle of old rhesus monkeys. *Mech. Ageing Dev.* **83**, 91–101 (1995).
- Khaidakov, M., Heflich, R. H., Manjanatha, M. G., Myers, M. B. & Aidoo, A. Accumulation of point mutations in mitochondrial DNA of aging mice. *Mutat. Res.* **526**, 1–7 (2003).
- Mueller-Hocker, J. Cytochrome-c-oxidase deficient cardiomyocytes in the human heart—An age-related phenomenon. *Am. J. Pathol.* **134**, 1167–1173 (1989).
- Fayet, G. *et al.* Ageing muscle: clonal expansions of mitochondrial DNA point mutations and deletions cause focal impairment of mitochondrial function. *Neuromuscul. Disord.* **12**, 484–493 (2002).
- Cottrell, D. A. *et al.* Cytochrome c oxidase deficient cells accumulate in the hippocampus and choroid plexus with age. *Neurobiol. Ageing* **22**, 265–272 (2001).
- Cottrell, D. A. & Turnbull, D. M. Mitochondria and ageing. *Curr. Opin. Clin. Nutr. Metab. Care* **3**, 473–478 (2000).
- Wallace, D. C. A mitochondrial paradigm for degenerative diseases and ageing. *Novartis Found. Symp.* **235**, 247–263 (2001).
- Carrodeguas, J. A., Kobayashi, R., Lim, S. E., Copeland, W. C. & Bogenhagen, D. F. The accessory subunit of *Xenopus laevis* mitochondrial DNA polymerase γ increases processivity of the catalytic subunit of human DNA polymerase γ and is related to class II aminoacyl-tRNA synthetases. *Mol. Cell. Biol.* **19**, 4039–4046 (1999).
- Foury, F. & Vanderstraeten, S. Yeast mitochondrial DNA mutators with deficient proofreading exonucleolytic activity. *EMBO J.* **11**, 2717–2726 (1992).
- Vanderstraeten, S., Van den Brule, S., Hu, J. & Foury, F. The role of 3'-5' exonucleolytic proofreading and mismatch repair in yeast mitochondrial DNA error avoidance. *J. Biol. Chem.* **273**, 23690–23697 (1998).
- Rodriguez, C. I. *et al.* High-efficiency deleter mice show that FLPe is an alternative to Cre-loxP. *Nature Genet.* **25**, 139–140 (2000).

- Meyers, E. N., Lewandoski, M. & Martin, G. R. An Fgf8 mutant allelic series generated by Cre- and Flp-mediated recombination. *Nature Genet.* **18**, 136–141 (1998).
- Kalu, D. N. *Handbook of Physiology* section 11 Aging (Oxford Univ. Press, New York, 1995).
- Arking, R. *Biology of Aging* (Sinauer, Sunderland, Massachusetts, 1998).
- Haines, D. C., Chattopadhyay, S. & Ward, J. M. Pathology of aging B6;129 mice. *Toxicol. Pathol.* **29**, 653–661 (2001).
- Weiss, A., Arbell, I., Steinhagen-Thiessen, E. & Silbermann, M. Structural changes in aging bone: osteopenia in the proximal femurs of female mice. *Bone* **12**, 165–172 (1991).
- Balducci, L. Epidemiology of anemia in the elderly: information on diagnostic evaluation. *J. Am. Geriatr. Soc.* **51**, S2–S9 (2003).
- Braunwald, E. (ed.) *Heart Disease. A Textbook of Cardiovascular Medicine* (W.B. Saunders Co., Philadelphia, 1997).
- Pal, L. & Santoro, N. Age-related decline in fertility. *Endocrinol. Metab. Clin. North Am.* **32**, 669–688 (2003).
- Tanemura, K., Kurohmaru, M., Kuramoto, K. & Hayashi, Y. Age-related morphological changes in the testis of the BDF1 mouse. *J. Vet. Med. Sci.* **55**, 703–710 (1993).
- Larsson, N. G. *et al.* Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. *Nature Genet.* **18**, 231–236 (1998).
- Silva, J. P. *et al.* Impaired insulin secretion and β -cell loss in tissue-specific knockout mice with mitochondrial diabetes. *Nature Genet.* **26**, 336–340 (2000).
- Sorensen, L. *et al.* Late-onset corticohippocampal neurodepletion attributable to catastrophic failure of oxidative phosphorylation in MILON mice. *J. Neurosci.* **21**, 8082–8090 (2001).
- Hansson, A. *et al.* A switch in metabolism precedes increased mitochondrial biogenesis in respiratory chain-deficient mouse hearts. *Proc. Natl Acad. Sci. USA* **101**, 3136–3141 (2004).
- Wredenberg, A. *et al.* Increased mitochondrial mass in mitochondrial myopathy mice. *Proc. Natl Acad. Sci. USA* **99**, 15066–15071 (2002).
- Wang, J., Silva, J., Gustafsson, C. M., Rustin, P. & Larsson, N. G. Increased *in vivo* apoptosis in cells lacking mitochondrial DNA gene expression. *Proc. Natl Acad. Sci. USA* **98**, 4038–4043 (2001).
- Spelbrink, J. N. *et al.* *In vivo* functional analysis of the human mitochondrial DNA polymerase POLG expressed in cultured human cells. *J. Biol. Chem.* **275**, 24818–24828 (2000).
- Bibb, M. J., VanEtten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. Sequence and organization of mouse mitochondrial DNA. *Cell* **26**, 167–180 (1981).
- Wibom, R., Hagenfeldt, L. & von Döbeln, U. Measurement of ATP production and respiratory chain enzyme activities in mitochondria isolated from small muscle biopsy samples. *Anal. Biochem.* **311**, 139–151 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements H.T.J., N.G.L. and J.T. are supported by the European Union MitAGE project. N.G.L. is supported by the Swedish Research Council, the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, the Torsten and Ragnar Söderbergs Foundation, the Swedish Heart and Lung Foundation, the Swedish Foundation for Strategic Research (Functional Genomics and INGVAR) and Funds of Karolinska Institutet. J.N.S. and H.T.J. are supported by the Academy of Finland and Tampere University Hospital Medical Research Fund. R.W. is supported by Funds of Karolinska Institutet and FreeMason's In Stockholm Foundation for Children's Welfare. A.O. is supported by the Swedish Research Council. We thank B. Rozell for mouse pathology evaluation, S. Hörtanainen, Z. Cansu, V. Edrisi and M. Åkerberg for technical assistance, and the Karolinska Center for Transgene Technologies for technical assistance.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to N.G.L. (nils-goran.larsson@mednub.ki.se).

An autonomous molecular computer for logical control of gene expression

Yaakov Benenson^{1,2}, Binyamin Gil², Uri Ben-Dor¹, Rivka Adar² & Ehud Shapiro^{1,2}

¹Department of Computer Science and Applied Mathematics and ²Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

Early biomolecular computer research focused on laboratory-scale, human-operated computers for complex computational problems^{1–7}. Recently, simple molecular-scale autonomous programmable computers were demonstrated^{8–15} allowing both input and output information to be in molecular form. Such computers, using biological molecules as input data and biologically active molecules as outputs, could produce a system for 'logical' control of biological processes. Here we describe an autonomous biomolecular computer that, at least *in vitro*, logically analyses the

levels of messenger RNA species, and in response produces a molecule capable of affecting levels of gene expression. The computer operates at a concentration of close to a trillion computers per microlitre and consists of three programmable modules: a computation module, that is, a stochastic molecular automaton^{12–17}; an input module, by which specific mRNA levels or point mutations regulate software molecule concentrations, and hence automaton transition probabilities; and an output module, capable of controlled release of a short single-stranded DNA molecule. This approach might be applied *in vivo* to biochemical sensing, genetic engineering and even medical diagnosis and treatment. As a proof of principle we programmed the computer to identify and analyse mRNA of disease-related genes^{18–22} associated with models of small-cell lung cancer and prostate cancer, and to produce a single-stranded DNA molecule modelled after an anticancer drug.

Taking our cue from the terminology of medical treatment, we consider that our molecular computer performs *in vitro* a computational version^{23,24} of ‘diagnosis’—the identification of a combination of mRNA molecules at specific levels, which in our example is a highly simplified model of cancer—and ‘therapy’—production of a biologically active molecule, which in our case is a drug-like single-stranded (ss)DNA with known anticancer activity (Fig. 1a). The computer operation is governed by a ‘diagnostic rule’ that encodes medical knowledge in simplified form (Fig. 1b). The left-hand side of the rule consists of a list of molecular indicators for a specific disease, and its right-hand side indicates a molecule to be released, which could be a drug for that disease. For example, the diagnostic rule for prostate cancer states²⁰ that if the genes *PPAP2B* and *GSTP1* are underexpressed and the genes *PIM1* and *HPN* are overexpressed then administer the ssDNA molecule GTTGGTATTGGACATG, which inhibits²⁵ the synthesis of the protein MDM2 by binding to its mRNA. The computer design is

flexible in that any sufficiently long RNA molecule can function as a molecular indicator and any short ssDNA molecule, up to at least 21 nucleotides, can be administered.

The computation module is a molecular automaton^{12–15} (Supplementary Fig. S1) that processes such a rule as depicted in Fig. 1c. The automaton has two states: positive (Yes) and negative (No). The computation starts in the positive state and if it ends in that state we call the result ‘positive diagnosis’, otherwise it is called ‘negative diagnosis’. To facilitate rule processing by the automaton, the left-hand side of the diagnostic rule is represented as a string of symbolic indicators, or symbols for short, one for each molecular indicator. For example, the string for the prostate cancer rule is *PPAP2B* ↓ *GSTP1* ↓ *PIM1* ↑ *HPN* ↑. For each symbol the automaton has three types of transitions: positive (Yes → Yes), negative (Yes → No) and neutral (No → No). The automaton processes the string from left to right, one symbol at a time. When processing a symbol in the positive state, the computer takes the positive transition if it determines that the molecular indicator is present and the negative transition, changing to a negative state, otherwise. As the No → Yes transition is not allowed, once the automaton enters the negative state it can use only the neutral transition and thus remains in the negative state for the duration of the computation. The possible computation paths of the automaton processing the prostate cancer diagnostic rule are shown in Fig. 1d.

The molecular automaton is stochastic^{14,16} in that it has two competing transitions, positive and negative, for each symbol while in the positive state. A novel molecular mechanism, explained below, regulates the probability of each positive transition by the corresponding molecular indicator, so that the presence of the indicator increases the probability of a positive transition and decreases the probability of its competing negative transition, and vice versa if the indicator is absent. Because the confidence with which the presence or absence of an indicator can be determined is a

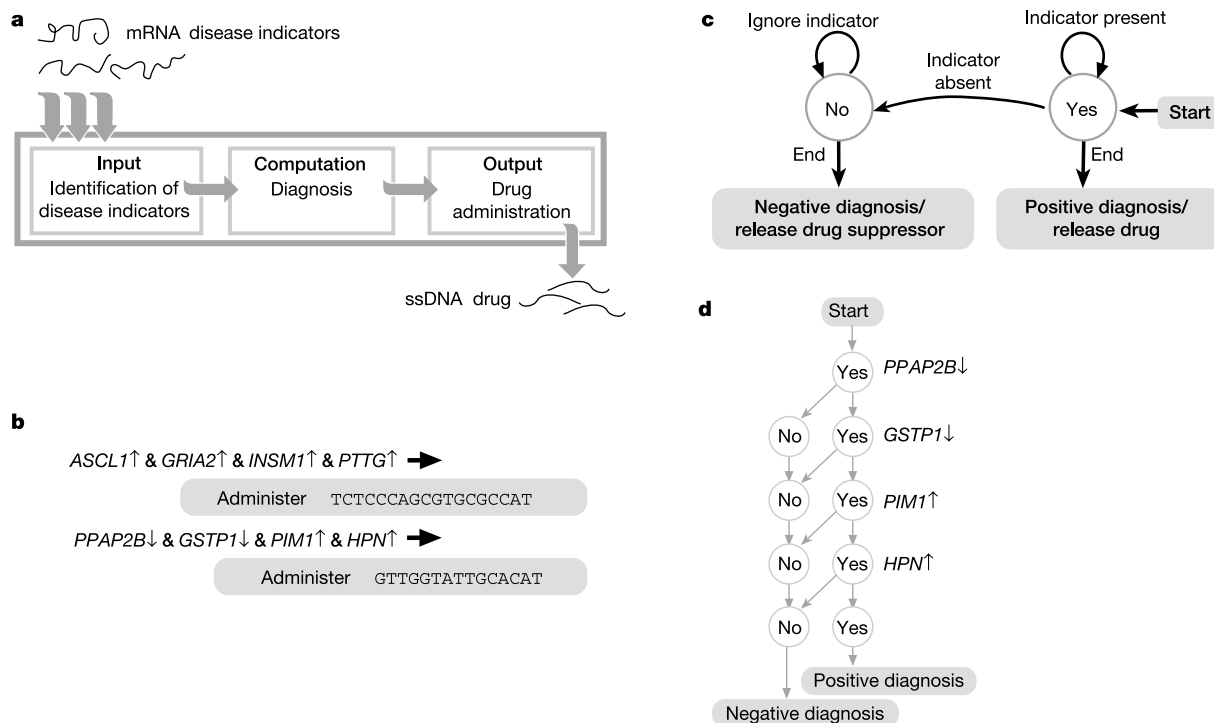


Figure 1 Logical design and logical operation of the molecular computer. **a**, Function and modular organization of the molecular computer. **b**, Example diagnostic rules for simplified models of SCLC¹⁹ and prostate cancer²⁰, indicating overexpression (↑) or underexpression (↓) of a disease-related gene. The first rule states that if genes *ASCL1*, *GRIA2*, *INSM1* and *PTTG1* are overexpressed then administer the ssDNA molecule

TCTCCAGCGTGCGCCAT (oblimersen), purported to be an antisense therapy drug for SCLC²⁷. The second rule states that if the genes *PPAP2B* and *GSTP1* are underexpressed and the genes *PIM1* and *HPN* are overexpressed then administer the ssDNA molecule GTTGGTATTGCACAT, purported to be a drug for prostate cancer²⁵. **c**, Transition diagram of the diagnostic automaton. **d**, The computation that diagnoses prostate cancer.

continuous rather than a discrete parameter, so is the regulation of transition probabilities, the level of which is correlated with this confidence. The resulting stochastic behaviour of the automaton is governed by the confidence in the presence of each indicator, so that

the probability of a positive diagnosis is the product of the probabilities of the positive transitions for each of the indicators processed (see Supplementary Note). By changing the ratio between positive and negative transitions for a particular indicator we can

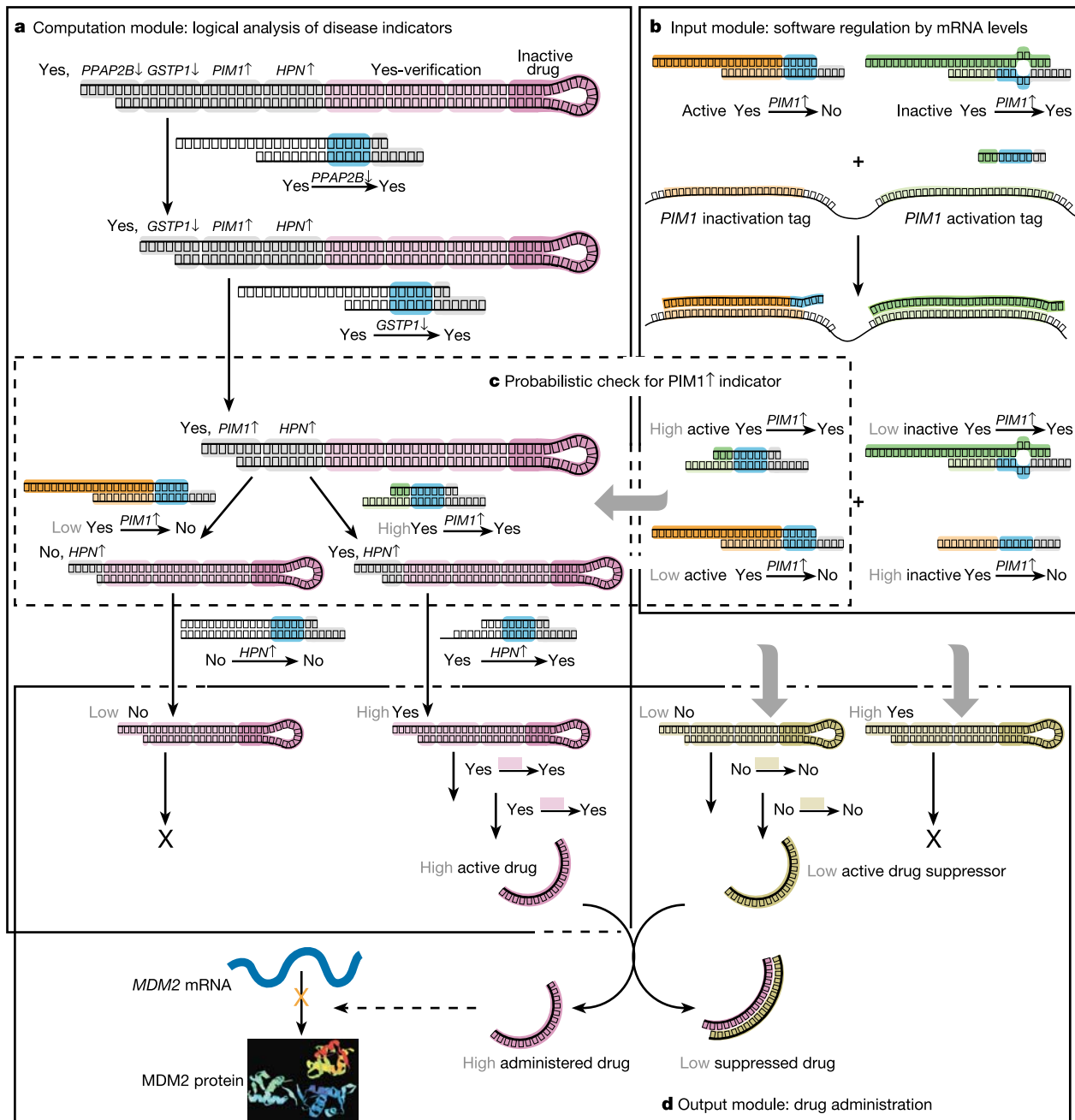


Figure 2 Operation of the molecular computer. The complete sequences for all molecules shown are given in the Supplementary Methods. **a**, Part of the computation path for the diagnostic molecule for prostate cancer with all molecular indicators present, ending in drug release. The initial diagnostic molecule consists of a diagnosis moiety (grey) that encodes the left-hand side of the diagnostic rule (Fig. 1b) and a drug-administration moiety (light purple) incorporating an inactive drug loop (dark purple). At each computation step, the prevailing transition is shown, except for the processing of the symbol $PIM1 \uparrow$, for which details of the stochastic choice, accomplished by a regulated pair of competing transition molecules, are shown (dashed box, see **c**). **b**, Regulation of the two transitions for $PIM1 \uparrow$ by sub-sequences (tags) of overexpressed $PIM1$ mRNA, resulting in a relatively high level of the $Yes \xrightarrow{PIM1 \uparrow} Yes$ transition molecules and low level of the $Yes \xrightarrow{PIM1 \uparrow} No$ molecules. Each transition molecule contains regulation (green, orange) and computation (blue, grey) fragments. The 'inactivation tag' of $PIM1$ mRNA (light

orange) displaces the $5' \rightarrow 3'$ strand of the transition molecule $Yes \xrightarrow{PIM1 \uparrow} No$ and destroys its computation fragment. The 'activation tag' of $PIM1$ mRNA (light green) activates the transition molecule $Yes \xrightarrow{PIM1 \uparrow} Yes$. Initially, a protecting oligonucleotide (green) partially hybridizes to the $3' \rightarrow 5'$ strand of the transition molecule and blocks the correct annealing of its $5' \rightarrow 3'$ strand. The 'activation tag' displaces the protecting strand, allowing such annealing and rendering an active $Yes \xrightarrow{PIM1 \uparrow} Yes$ transition. Ideally, one $PIM1$ mRNA molecule inactivates one $Yes \xrightarrow{PIM1 \uparrow} No$ and activates one $Yes \xrightarrow{PIM1 \uparrow} Yes$ transition molecule. **c**, Stochastic processing of the symbol $PIM1 \uparrow$ by a regulated pair of competing transition molecules. The probability of a $Yes \rightarrow Yes$ transition is high, resulting in a high level of diagnostic molecules in the state Yes and a low level in state No . **d**, Combining computation results for both types of diagnostic molecules, both with high Yes and low No final states, results in high release of drug and low release of drug suppressor, and hence in the administration of the drug.

have fine control over the sensitivity of diagnosis to the presence of that indicator. We note our unusual use of automaton components: its formal input (the diagnostic rule to be processed) functions in our application like a program, and its formal program (the software molecules) functions in our application as part of the input module, detecting the presence of molecular indicators.

Instead of releasing an output molecule on positive diagnosis and doing nothing on negative diagnosis, we opted to release a biologically active molecule (for example, a drug) on positive diagnosis and its suppressor molecule on negative diagnosis. This allows fine control over the diagnosis confidence threshold beyond which an active drug is administered. Rather than using a single automaton

for both tasks (which cannot be done with our current design), we implement this by using two types of automata: one that releases a drug molecule on positive diagnosis, and another that releases a drug-suppressor molecule on negative diagnosis. The ratio between the drug and drug-suppressor molecules released by a population of automata of these two types determines the final active drug concentration.

The molecular computer realizing this logical design consists of diagnostic molecules that encode diagnostic rules (Fig. 2a; see also Supplementary Fig. S2); transition molecules that realize automaton transitions and are regulated by molecular indicators (Fig. 2b; see also Supplementary Fig. S2); and hardware molecules (the

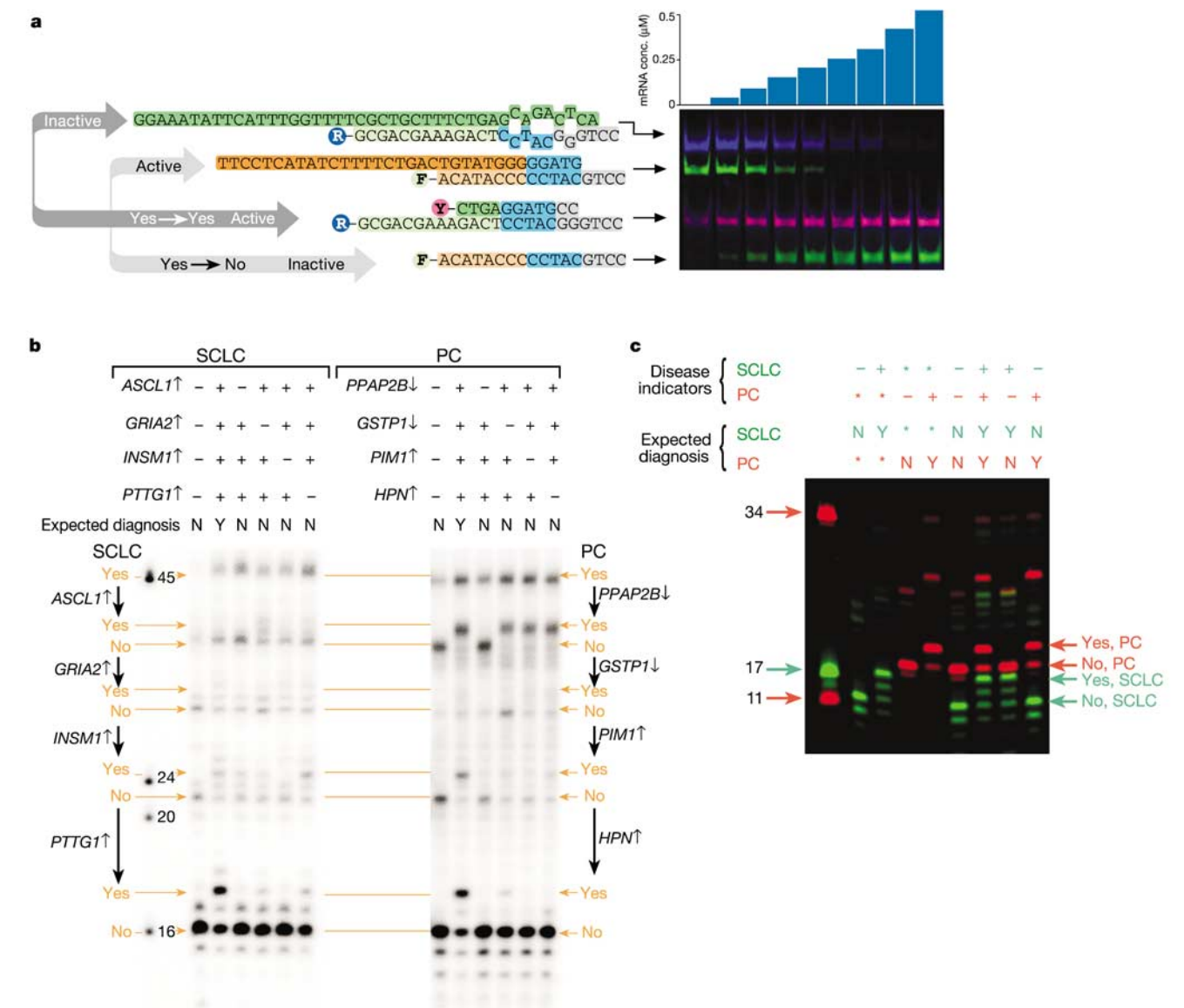


Figure 3 Experimental demonstration of diagnosis. **a**, Regulation of competing transitions by pTRI-Xef mRNA (bar chart) representing an example molecular indicator. The gel shows positive (blue and pink) and negative (green) transition, each in both active and inactive states, the relative concentration of which is regulated by mRNA concentration. F, carboxyfluorescein; R, tetramethyl rhodamine; Y, Cy5. **b**, Validation of the diagnostic automata with the diagnostic rules for SCLC and prostate cancer (PC). Each lane shows the result of a diagnostic computation for the indicated combination of molecular indicators. The location of the initial, intermediate and output molecules is indicated on

the left and a predicted trace of the diagnostic computations is shown on both sides. Some, but not all, intermediates are visible owing to their incomplete processing by the automaton. **c**, Parallel diagnosis of two disease models by two diagnostic automata. The diagnosed disease model contains the two indicators of SCLC checked by the diagnostic string *PTTG1* ↑ *CDKN2A* ↑ and the two indicators of prostate cancer checked by the string *PIM1* ↑ *HPN* ↑. The gene *CDKN2A* is another indicator for SCLC when overexpressed, and it is used here for technical reasons. The presence of indicators for each disease model as well as the expected diagnostic output by each automaton are indicated above the lanes.

restriction enzyme *FokI* (Supplementary Fig. S1).

A diagnostic molecule has a diagnosis moiety and a drug-administration moiety (Fig. 2a; see also Supplementary Fig. S2). The diagnosis moiety realizes each symbol by a unique double-stranded (ds)DNA sequence 7 base pairs (bp) in length. For all symbol-encoding sequences, the first four nucleotides represent the symbol combined with the positive state and nucleotides three to six represent the symbol combined with the negative state. The automaton processes the diagnosis moiety one symbol at a time, determining whether the corresponding indicator is present (Fig. 1c). After all symbols are processed, the computation proceeds to processing the drug-administration moiety. There are two kinds of drug-administration moieties: drug release and drug-suppressor release. Both consist of a double-stranded stem that protects a single-stranded loop containing the drug or the drug suppressor. The stem prevents interactions between the drug, drug suppressor and target mRNA. Its length (three symbols) was determined empirically (data not shown). After positive diagnosis, the stem of a drug-suppressor moiety is cleaved by special Yes-verification transitions and releases the drug; in the case of negative diagnosis it remains intact, protecting the drug (Fig. 2d). After negative diagnosis, the stem of the drug-suppressor release moiety is cleaved

by special No-verification transitions and releases the drug suppressor; in the case of positive diagnosis it remains intact, protecting the drug suppressor (Fig. 2d).

The ratio between diagnostic molecules with a drug release moiety and those with a drug-suppressor release moiety need not necessarily be 1:1. For example, if the measured probability of positive diagnosis does not exceed 50% owing to computer design limitations, combining the two types of molecules at a ratio of 2:1 will render an active drug for 1/6 of the computations. Inverting this ratio can compensate for false-positive diagnosis by suppressing drug release below any desired threshold.

The molecules released in our experiments are the antisense ssDNA for MDM2 (GTTGGTATTGCACAT)²⁵ and the ssDNA molecule having the same sequence as the drug Vitravene²⁶ (GCGTTTGCTCTTCTTCTTGCG) (Supplementary Data). Similar molecules, such as oblimersen²⁷ (TCTCCAGCGTGCGCCAT), could similarly be released. The same method can be used to release short RNA molecules.

For each symbolic indicator, a pair of competing transition molecules (Fig. 2b; see also Supplementary Fig. S2) perform the corresponding molecular indicator verification. Presence of a molecular indicator entails high concentration of the positive transition

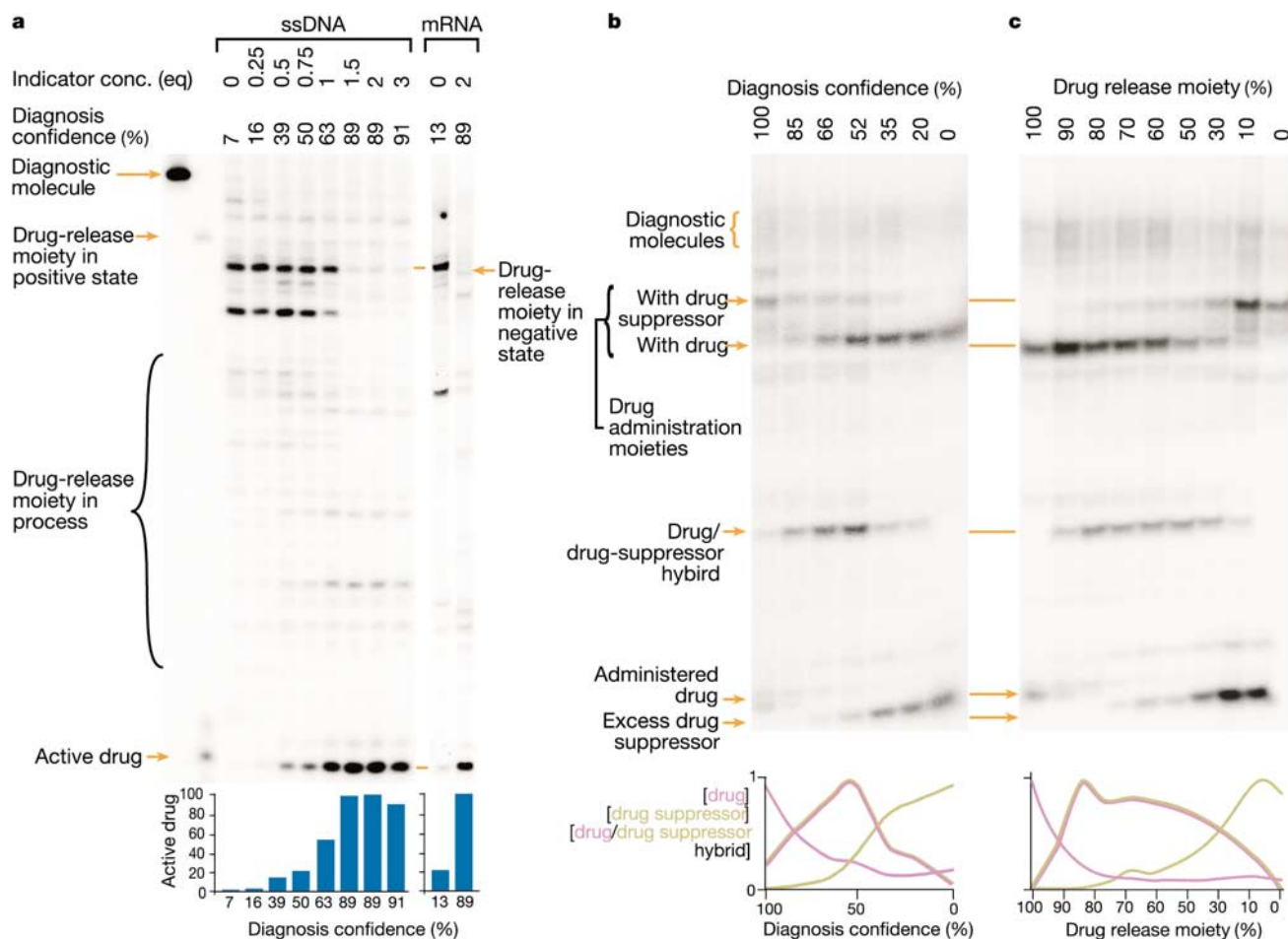


Figure 4 Experimental demonstration of drug administration. **a**, Release of an active drug by a drug-release *PPAP2B* ↓ *GSTP1* ↓ *PIM1* ↑ *HPN* ↑ diagnostic molecule, showing absolute amount of the active drug versus positive diagnosis probability. **b**, Different diagnostic outcomes are modelled using active transition molecules with a mixture of equal amounts of the drug-release and drug-suppressor release moieties for the diagnostic string *PPAP2B* ↓ *GSTP5* ↓. Each lane shows the distribution of drug-

administration moieties, active drug, excess drug suppressor and drug/drug-suppressor hybrid, as indicated. **c**, Variation in the distribution of active drug, excess drug suppressor and drug/drug-suppressor hybrid for a given diagnostic outcome and for varying relative amounts of drug release and drug-suppressor release diagnostic moiety. y-axis units in **b** and **c** are arbitrary units.

molecule and low concentration of its competing negative transition molecule and vice versa. This regulation is accomplished by sequence-specific interaction between the indicator and a partially single-stranded fragment of a transition molecule, as follows. A positive transition checking for overexpression is activated by a high concentration of its corresponding mRNA, whereas a positive transition checking for underexpression is inactivated by a high concentration of its corresponding mRNA. The corresponding negative transitions are oppositely regulated by a similar mechanism (Fig. 2b; see also Supplementary Fig. S2). A similar mechanism allows for transition regulation by point mutation (Supplementary Fig. S2). The logical switch between configurations of the transition molecules is similar to the state that switching a DNA fuel molecule causes in a DNA nanoactuator²⁸. We also note an alternative approach to sensing biochemical signals, known as 'chemical logic gates'^{29,30}.

The computer consists of three molecular modules—input (Fig. 2b), computation (Fig. 2a) and output (Fig. 2d)—that interact with the disease-related molecules and with each other via a complex network (Fig. 2). Each molecular computer autonomously performs one diagnosis and drug administration task; multiple tasks can be performed by multiple computers that operate in parallel within the same environment without mutual interference, while sharing the hardware molecules and potentially sharing some or all of the software molecules. All pairs of transition molecules are regulated simultaneously by their respective indicators (Fig. 2b) and perform a stochastic computation over diagnostic molecules to administer a drug upon diagnosis (Fig. 2d).

To provide a successful implementation, the computer must be robust both to imprecision of molecular components and to variations in external parameters. This is achieved by three mechanisms. First, imprecision in transition regulation may be compensated by variation in the relative amounts of the active and inactive transition molecules and by addition of excess ssDNA oligonucleotides that form these transitions. Second, changes in the absolute level at which a molecular indicator should be positively detected are compensated for by a similar change in absolute concentration of the transition molecules (Supplementary Fig. S3). Third, false-positive or false-negative diagnoses may be compensated for as explained above.

The operation of the computer modules was verified separately and jointly: transition regulation by the input module (Fig. 2b; see also Supplementary Fig. S2) was verified independently (Fig. 3a; see also Supplementary Fig. S3); the input and computation modules performing transition regulation and diagnostic computation (Fig. 2a) were verified together (Fig. 3b, c); finally, the input, computation and output modules were combined to perform transition regulation, diagnostic computation and drug administration (Figs 2d and 4a).

The regulation of transition molecules by mRNA (Fig. 2b) was demonstrated (Fig. 3a) with an mRNA transcript of about 1,900 nucleotides as an example indicator. A correlation between the mRNA level and the probability of positive diagnosis, effected by a pair of transitions regulated to check for underexpression, is shown in Supplementary Fig. S3. Regulation by mutation was confirmed with a ssDNA model simulating a point substitution representing a small-cell lung cancer (SCLC)-related mutation in the gene encoding p53 (ref. 21) (Supplementary Fig. S3). We expect to extend this approach to detect insertion and deletion mutations.

In diagnosing a disease model with multiple indicators we used ssDNA oligonucleotides to represent disease-related mRNA and two concentrations to represent mRNA levels: 0 μ M for low level and 3 μ M for high level. Transition regulation can be adjusted by changing the absolute concentration of competing transitions to distinguish between no mRNA and mRNA at concentrations as low as 100 nM, which represent about 50 mRNA copies per mammalian

cell. In our experiments we used up to four molecular indicators, although the specific symbol encoding used can provide up to eight indicators.

We tested the input and computation modules of the computer on molecular models of SCLC and prostate cancer with diagnostic automata for the diagnostic rules shown in Fig. 1b. The diagnostic output was identified by the size of an inert fragment released on completion of the computation. Each automaton reaches positive diagnosis with significant probability only when all four molecular indicators are present (Fig. 3b). The false-negative diagnosis is due to limitations in the design of the transition molecules, but can be compensated for during drug administration (Fig. 4c) as discussed above. We also confirmed the selectivity of the diagnostic process by testing the two diagnostic automata in mixed conditions, in which molecular indicators for none, one, or both disease models are present (Supplementary Fig. S4). In all cases a positive diagnosis was made with significant probability by a diagnostic automaton only when all molecular indicators were present. We confirmed the possibility of parallel diagnosis of multiple disease models by running together two diagnostic automata under various compositions of molecular indicators (Fig. 3c). Indeed, each automaton performed its diagnosis correctly, irrespective of the computation performed by the other automaton.

Drug administration is demonstrated for the prostate cancer disease model (Fig. 4). We constructed the drug-release diagnostic molecule for *PPAP2B* $\downarrow$ *GSTP1* $\downarrow$ *PIM1* $\uparrow$ *HPN* $\uparrow$ (Fig. 2a) and tested the extent of active drug release for different diagnostic outcomes, effected by varying amounts of ssDNA representing *HPN* mRNA and, in a separate experiment, an example mRNA that substitutes for *GSTP1* mRNA. The presence of other indicators was modelled by appropriately formed positive transitions. We show that the amount of active drug increases with the confidence in a positive diagnosis (Fig. 4a). We demonstrate the concept of drug regulation by a drug suppressor using diagnostic molecules for *PPAP2B* $\downarrow$ *GSTP1* $\downarrow$ with drug release and drug-suppressor release moieties. We show that the prevailing species is the active drug for high probability values, a drug/drug-suppressor hybrid for intermediate probability values, and an active drug suppressor for low probability values (Fig. 4b). We also demonstrate the ability to control the relative amounts of drug and drug suppressor for the 1:1 ratio of positive and negative diagnosis (Fig. 4c). The result demonstrates the robustness of our proposed compensation mechanism and illustrates how multiple degrees of freedom of the system allow it to overcome imperfections in its components.

Our work demonstrates a robust and flexible molecular computer capable of logical analysis of mRNA disease indicators *in vitro* and the controlled administration of biologically active ssDNA molecules, including drugs. The modularity of the design facilitates improving each computer component independently. For example, computer regulation by other biological molecules such as proteins, the output of other biologically active molecules such as RNA interference, and *in vivo* operation can all be explored simultaneously and independently. $\square$

Received 4 February; accepted 6 April 2004; doi:10.1038/nature02551.

Published online 28 April 2004.

- Adelman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994).
- Lipton, R. J. DNA solution of hard computational problem. *Science* **268**, 542–545 (1995).
- Ouyang, Q., Kaplan, P. D., Liu, S. & Libchaber, A. DNA solution of the maximal clique problem. *Science* **278**, 446–449 (1997).
- Khodor, J. & Gifford, D. K. Design and implementation of computational systems based on programmed mutagenesis. *Biosystems* **52**, 93–97 (1999).
- Faulhammer, D., Cukras, A. R., Lipton, R. J. & Landweber, L. F. Molecular computation: RNA solutions to chess problems. *Proc. Natl Acad. Sci. USA* **97**, 1385–1389 (2000).
- Liu, Q. *et al.* DNA computing on surfaces. *Nature* **403**, 175–179 (2000).
- Ruben, A. J. & Landweber, L. F. The past, present and future of molecular computing. *Nature Rev. Mol. Cell Biol.* **1**, 69–72 (2000).
- Winfree, E., Liu, F. R., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998).

9. Mao, C., LaBean, T. H., Reif, J. H. & Seeman, N. C. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature* **407**, 493–496 (2000).
10. Sakamoto, K. *et al.* State transitions by molecules. *Biosystems* **52**, 81–91 (1999).
11. Sakamoto, K. *et al.* Molecular computation by DNA hairpin formation. *Science* **288**, 1223–1226 (2000).
12. Benenson, Y. *et al.* Programmable and autonomous computing machine made of biomolecules. *Nature* **414**, 430–434 (2001).
13. Benenson, Y., Adar, R., Paz-Elizur, T., Livneh, Z. & Shapiro, E. DNA molecule provides a computing machine with both data and fuel. *Proc. Natl Acad. Sci. USA* **100**, 2191–2196 (2003).
14. Adar, R. *et al.* Stochastic computing with biomolecular automata. *Proc. Natl Acad. Sci. USA* (submitted).
15. Benenson, Y. & Shapiro, E. in *Dekker Encyclopedia of Nanoscience and Nanotechnology* (eds Schwarz, J. A., Contescu, C. I. & Putyera, K.) 2043–2056 (Dekker, New York, 2004).
16. Rabin, M. O. Probabilistic automata. *Inform. Contr.* **6**, 230–245 (1963).
17. Bar-Ziv, R., Tlusty, T. & Libchaber, A. Protein-DNA computation by stochastic assembly cascade. *Proc. Natl Acad. Sci. USA* **99**, 11589–11592 (2002).
18. Sidransky, D. Emerging molecular markers of cancer. *Nature Rev. Cancer* **2**, 210–219 (2002).
19. Pedersen, N. *et al.* Transcriptional gene expression profiling of small cell lung cancer cells. *Cancer Res.* **63**, 1943–1953 (2003).
20. Dhanasekaran, S. M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
21. Takahashi, T. *et al.* The p53 gene is very frequently mutated in small-cell lung cancer with a distinct nucleotide substitution pattern. *Oncogene* **6**, 1775–1778 (1991).
22. Thorns, C., Gaiser, T., Lange, K., Metz, H. & Feller, A. C. cDNA arrays: Gene expression profiles of Hodgkin's disease and anaplastic large cell lymphoma cell lines. *Pathol. Int.* **52**, 578–585 (2002).
23. Ledley, R. S. & Lusted, L. B. Reasoning foundation of medical diagnosis. *Science* **130**, 9–21 (1959).
24. Holzer, S., Fremgen, A. M., Hundahl, S. A. & Dudeck, J. Analysis of medical-decision making and the use of standards of care in oncology. *J. Am. Med. Assoc. (Suppl. S)* 364–368 (2000).
25. Capoulade, C. *et al.* Apoptosis of tumoral and nontumoral lymphoid cells is induced by both mdm2 and p53 antisense oligodeoxynucleotides. *Blood* **97**, 1043–1049 (2001).
26. Holmlund, J. T. Applying antisense technology. *Ann. NY Acad. Sci.* **1002**, 244–251 (2003).
27. Klasa, R. J., Gillum, A. M., Klem, R. E. & Frankel, S. R. Oblimersen Bcl-2 antisense: Facilitating apoptosis in anticancer treatment. *Antisense Nucleic Acid Drug Dev.* **12**, 193–213 (2002).
28. Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605–608 (2000).
29. Stojanovic, M. N. & Stefanovic, D. A deoxyribozyme-based molecular automaton. *Nature Biotechnol.* **21**, 1069–1074 (2003).
30. Balzani, V., Credi, A. & Venturi, M. Molecular logic circuits. *Chemphyschem* **4**, 49–59 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Katrav for the design and preparation of the figures; G. Linshiz for discussion and help in oligonucleotide purification; Z. Livneh for encouraging us to pursue this research direction; A. Regev for critical review and suggestions; and M. Vardi for discussion and references. This work was supported by the Moross Institute for Cancer Research, Israeli Science Foundation and the Minerva Foundation.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to E.S. (Ehud.Shapiro@weizmann.ac.il).

Assembly and function of a bacterial genotoxin

Dragana Nešić, Yun Hsu & C. Erec Stebbins

Laboratory of Structural Microbiology, The Rockefeller University, New York 10021, USA

The tripartite cytolethal distending toxin (CDT) induces cell cycle arrest and apoptosis in eukaryotic cells^{1,2}. The subunits CdtA and CdtC associate with the nuclease CdtB to form a holotoxin that translocates CdtB into the host cell, where it acts as a genotoxin by creating DNA lesions^{3–7}. Here we show that the crystal structure of the holotoxin from *Haemophilus ducreyi* reveals that CDT consists of an enzyme of the DNase-I family, bound to two ricin-like lectin domains. CdtA, CdtB and CdtC form a ternary complex with three interdependent molecular interfaces, characterized by globular, as well as extensive non-globular, interactions. The lectin subunits form a deeply

grooved, highly aromatic surface that we show to be critical for toxicity. The holotoxin possesses a steric block of the CdtB active site by means of a non-globular extension of the CdtC subunit, and we identify putative DNA binding residues in CdtB that are essential for toxin activity.

Many Gram-negative pathogens, such as *Escherichia coli*, *Shigella dysenteriae*, *Actinobacillus actinomycetemcomitans*, *Campylobacter* spp., *Helicobacter* spp., *Salmonella typhi* and *H. ducreyi*, possess a related toxic activity that causes cell cycle arrest and subsequent cellular distension in epithelial cells, and a rapid death by apoptosis in many lymphocytes^{2–4,8}. This toxic activity is associated with the products of an operon that encodes three proteins: CdtA, CdtB and CdtC (refs 2–4). CdtB shares conserved residues with the active sites of DNase I-like nucleases^{9,10} and demonstrates nuclease activity *in vitro* and *in vivo*^{7,8,10–16}. CdtA and CdtC are required to form a tripartite complex with CdtB that delivers the CdtB subunit into cells, leading many eukaryotic cell types to activate DNA damage checkpoints^{3,5–8,14–21}. Taken together, these facts suggest that this toxin may have carcinogenic properties^{3,4,10}.

The crystal structure of the holotoxin from *H. ducreyi* reveals that CDT is a ternary complex with three extensive globular protein–protein interfaces (burying a total of over 6,300 Å² of surface area), between CdtA–CdtB, CdtA–CdtC and CdtB–CdtC (Fig. 1a). In addition, CdtA and CdtC possess amino and carboxy termini that project as non-globular polypeptides (Figs 1 and 2) and interact extensively with each other and with CdtB (altogether burying another 4,400 Å² of surface area), thereby contributing to an overall massive and interdependent surface area of interaction (over 10,700 Å² for the entire holotoxin, with relative molecular mass 70,000 (M_r 70K)).

CdtA and CdtC are both lectin-type structures, homologous to the B-chain repeats of the plant toxin, ricin²², consisting of a pseudo three-fold symmetric trefoil of 12 β-strands, arranged in three separate β-sheets. CdtA and CdtC align to the first repeat of the ricin B-chain with root mean square (r.m.s.) deviations of 2.3 Å and 3.7 Å over 119 and 108 amino acids, respectively (with 16% sequence identity between ricin and CdtA from a structural alignment)²³. CdtA and CdtC bind at one end of CdtB, contacting each other (Fig. 1) in a manner reminiscent of the repeat arrangement in the ricin B-chain and its interaction with the active enzymatic subunit of this toxin²², although CdtA and CdtC cannot be superimposed as a rigid unit with the ricin repeats.

CdtA and CdtC together contribute towards the formation of two notable surface elements, both located on the face of the holotoxin that is directed away from CdtB. The first is a large aromatic cluster consisting of eight bulky side-chains, most of which are well conserved in CDT homologues (Fig. 1b). The second is a long and deep groove, or crevice, (approximately 20 × 10 Å long and 5–8 Å deep) that is formed from structural components of both proteins (Figs 1c and 2a).

Mutagenesis of the aromatic patch renders the holotoxin inactive in cellular assays, although the ternary complex stability is unimpaired (Fig. 3). Previous reports have shown that CdtA and CdtC are each able to bind to the surface of cells^{24–26}, although the data for CdtC are conflicting¹². Our structure and mutagenesis strongly support the idea that CdtA (through the aromatic patch) has an important role in the activity of this toxin, presumably through cell surface binding. Because CdtA and CdtC both contribute to the formation of the deep groove, the function of this structural element might be severely impaired when one of the components is absent, and may explain why full toxicity is only achieved when both components are together^{17,24}. The CdtA N-terminal extension (discussed below) forms its crystal-packing contacts by interacting with this groove and inserting itself deeply into it (Fig. 2a). This raises the possibility that the CdtA N terminus in the crystals may be mimicking a peptide from a cell surface receptor for CDT, implicating the groove as a potential peptide-binding cleft.

The close interplay of CdtA and CdtC in the formation of the groove and aromatic patch, and the similarity in their positioning with the two lectin repeats in the ricin B-chain, suggest that these two components of CDT work together for a related function. This function is likely to be receptor binding to initiate endocytosis of the complex, and strongly supports the hypothesis that the CdtA–CdtC sub-complex can be viewed as the ‘B’ or ‘binding’ element of a two-component AB toxin, as had been proposed from previous biochemical studies⁵. CDT would therefore be an AB₂ toxin, with CdtB taking the role of the active, or ‘A’, subunit.

CdtA and CdtC each have extended, non-globular polypeptides at their N and C termini, and in both proteins these regions interact with other elements of the holotoxin to cement the ternary complex together. The N terminus of the crystallized construct of CdtA extends to residue 18, and amino acids 18–56 are disordered. The non-globular extension from the CdtA N terminus is highly unusual, as amino acids 56–67 make only minor contacts to any elements of the holotoxin in the crystal asymmetric unit; instead they extend distally to make extensive crystal-packing contacts with a symmetry-related complex (Figs 1 and 2a). This is likely to be a crystal-packing artefact, as the CDT complex shows no signs of oligomerization by gel-filtration chromatography (Fig. 3a). It is probable, therefore, that residues 18–67 are normally disordered in the complex, but a fortuitous crystal-packing arrangement allowed for the region between 56 and 67 to become ordered. The amino acids 67–75 of this extended polypeptide, however, do contact CdtB and CdtC in the appropriate complex in the crystal, and therefore represent a true structural interaction. *H. ducreyi* CdtA is smaller than CdtA proteins from most other bacterial species, and the additional polypeptide is likely to be incorporated into the N-terminal region, as judged by sequence alignment and sequence threading to our crystal structure (data not shown).

The N-terminal 13 amino acids of CdtC extend away from the globular β -trefoil domain and interact with CdtB as a non-globular

polypeptide that covers the active site of the nuclease (Figs 1a and 2b), interacting with active site residues, such as His 160, and hypothesized DNA binding residues like Arg 117 and Arg 144 of CdtB (discussed below). These contacts make the holotoxin, as found in the crystals, completely incompatible with DNA binding to the active site of CdtB (Fig. 2b). This feature of the structure raises the possibility that the CdtC N-terminal ‘tail’ inhibits CdtB with, perhaps, an auto-inhibitory function for the holotoxin. A mutant that deletes the N-terminal residues that occlude the CdtB active site, CdtC Δ (21–35), was assembled into the holotoxin, and this complex was examined in toxicity and nuclease assays (Figs 3b and 4c). CdtC Δ (21–35) assembles into a biochemically stable holotoxin (Fig. 3a) that possesses activity against plasmid DNA similar to that of CdtB alone (Fig. 4c). The wild-type holotoxin, however, showed no activity in this plasmid-relaxation assay (Fig. 4c). Interestingly, this mutant holotoxin is as active as the wild-type complex in cellular toxicity assays (Fig. 3b). Therefore, the N-terminal tail of CdtC appears to inhibit the nuclease activity of the assembled holotoxin *in vitro*, while not serving a role in cellular intoxication.

The active subunit of the CDT holotoxin is the CdtB protein, which is the most highly conserved of the three holotoxin components and shows weak sequence similarity to proteins of the DNase-I family (approximately 12% identity based on a structural alignment)²³. The crystal structure reveals that CdtB adopts the canonical, four-layered fold for this family of enzymes, consisting of a central 12-stranded β -sandwich packed between outer α -helices and loops on each side of the sandwich. DNase I, HAP1 (human DNA repair endonuclease), and ExoIII (exonuclease III of *E. coli*) align with CdtB over the core fold with r.m.s. deviations in C α positions of 3.3 Å, 3.3 Å and 2.9 Å, over 209, 198 and 197 residues, respectively. Overall, CdtB is most structurally similar to the DNase-I fold (as judged by the Z score²³).

In common with DNase I, CdtB possesses two absolutely conserved histidine residues (CdtB His 160 and His 274—Fig. 4a). These

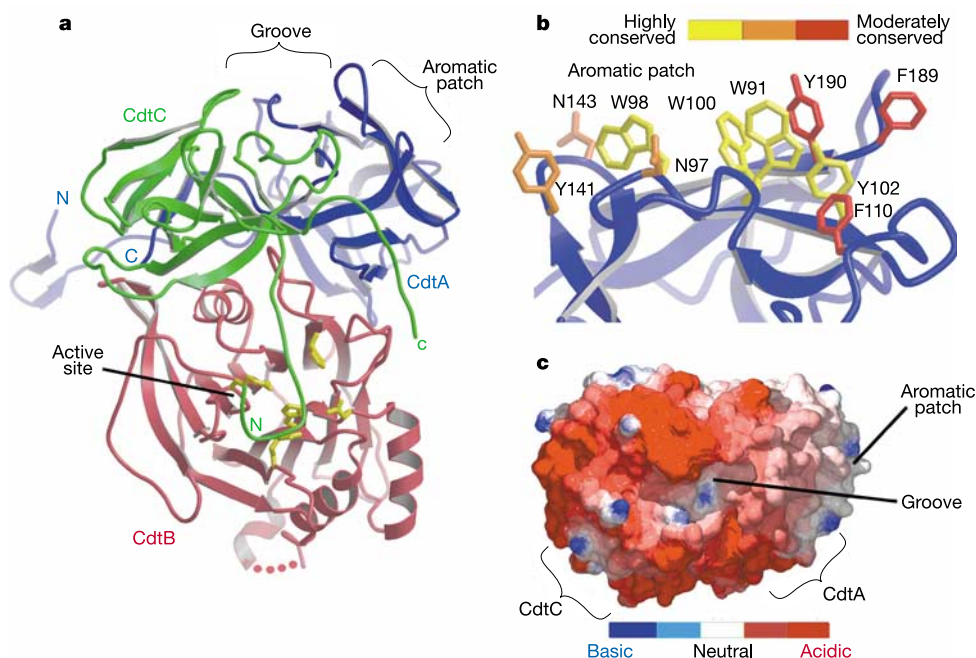


Figure 1 Assembly and features of the heterotrimeric cytolethal distending toxin.

a, Crystal structure of the *H. ducreyi* CDT, shown as a ribbon cartoon tracing the three polypeptide chains. The active site and possible DNA contacting residues of CdtB are shown in yellow. Red dots indicate CdtB peptide not modelled owing to disorder. N, amino terminus; C, carboxy terminus. **b**, Aromatic side chains clustered on the surface of CdtA

are coloured by degree of conservation between CdtA homologues, as indicated. The patch is at the top of the molecule, and the orientation of the figure is similar to CdtA in **(a)**, rotated by 180° about a vertical axis. **c**, Surface groove and electrostatics of the CdtA–CdtC ‘top’ portion of the holotoxin (CdtB is beneath or ‘behind’ the surface shown, and the orientation of the holotoxin is like that in **(a)**, rotated 90° about a horizontal axis).

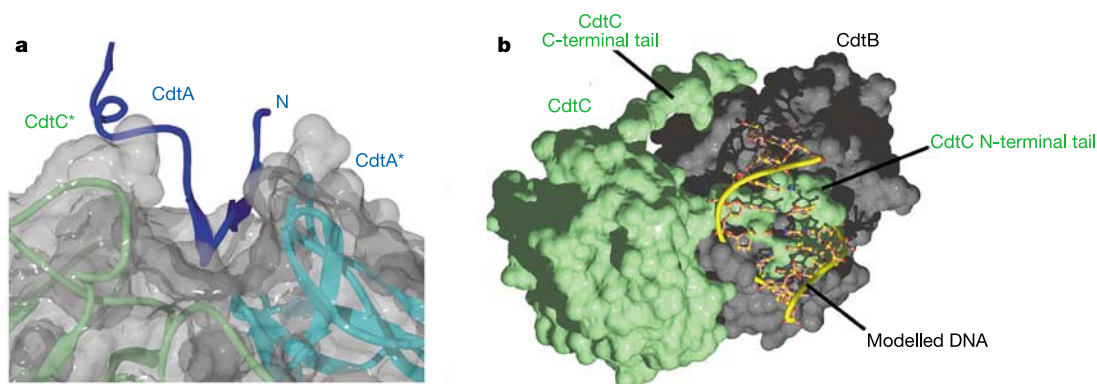


Figure 2 Two non-globular extensions in the holotoxin. **a**, Surface representation (grey) of the crystallographic symmetry-related interaction between the groove and the N-terminal tail of CdtA. The surface is in a similar orientation to Fig. 1a, and is shown partially transparent. The symmetry-related holotoxin subunits are denoted by an asterisk. **b**, The

molecular surfaces of CdtB and CdtC are shown in grey and green, respectively. DNA, positioned by aligning the DNA–DNase I complex with CdtB, is drawn with the phosphodiester backbone in yellow and the bases in light red, with the atoms of carbon, nitrogen and oxygen coloured yellow, blue and red, respectively.

putative catalytic histidines of CdtB are critical for toxin activity^{9,10} and superimpose almost perfectly with the active site histidines in structural alignments with DNase I (Fig. 4a). Only one of the catalytic histidines (CdtB His 274) hydrogen bonds with a nearby carboxylate group (CdtB Asp 238), as is observed in DNase I; an interaction that has been postulated to be important due to its raising of the histidine pK_a value during phosphodiester cleavage and for correctly orienting the catalytic residues^{27–29}. His 160 (equivalent to DNase I His 134) does not have such a bonding partner, because Val 118 of CdtB occupies the equivalent position of DNase I Glu 78 (Fig. 4a).

Because mutation of these partner aspartic or glutamic acid residues in DNase I significantly lowers the activity of this enzyme^{29,30}, it could be hypothesized that the very low DNase activity observed for CdtB *in vitro* (0.01% the activity of bovine DNase I)^{11,12} may be due to the presence of a valine (Val 118),

instead of glutamate, at this position. *In vitro*, plasmid relaxation assays of wild type and a Val118Glu mutant, however, show that the mutant is inactive in the conversion of supercoiled to relaxed or linear forms (Fig. 4c). Supporting the *in vitro* results, the Val118Glu mutation rendered the assembled holotoxin completely inactive in a cellular toxicity assay (Fig. 4d), despite the fact that the ternary complex assembled *in vitro* indistinguishably from the wild type (data not shown). These results indicate that this apparently simple reversion (to glutamate) cannot be tolerated in CdtB.

Three highly conserved residues in DNase I, that make direct contacts to substrate DNA, have nearly identical counterparts in CdtB (Fig. 4a, b). Arg 111 and Asn 170 of DNase I have similar three-dimensional positions to CdtB Arg 144 and Asn 201 (two absolutely conserved residues in the CdtB subunit). A third DNA contact of DNase I, Arg 41, also has a three-dimensionally equivalent residue from CdtB, Arg 117. Although the C α positions of

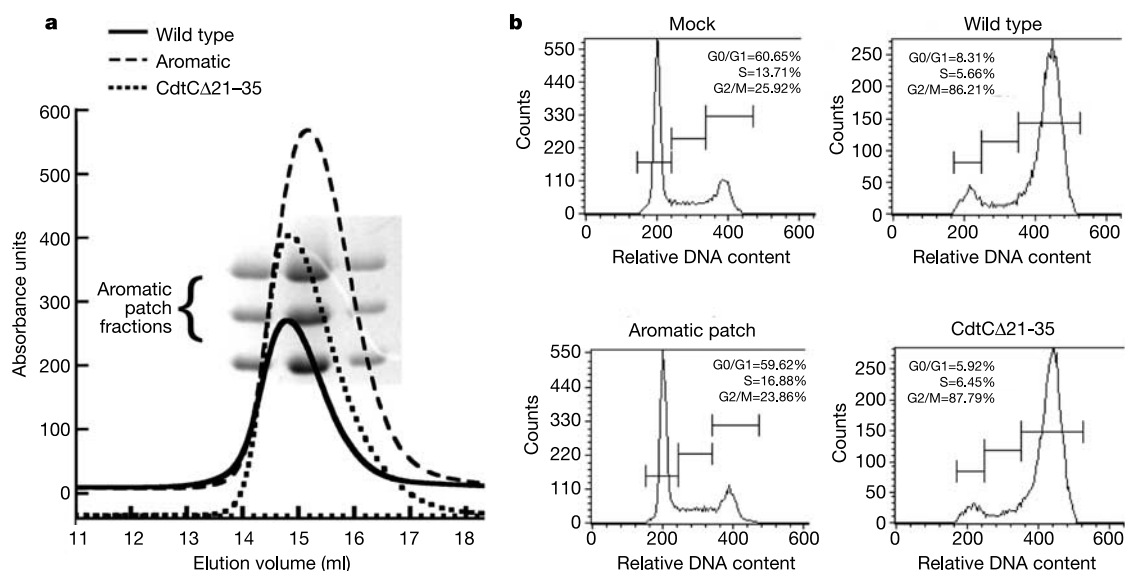


Figure 3 Structure-based mutagenesis of the lectin subunits. **a**, *In vitro* size exclusion (gel filtration) chromatography of different CdtA and CdtC mutants was used to assay holotoxin integrity. The corresponding fractions of the CdtA aromatic patch mutant (quadruple Trp91Gly, Trp98Gly, Trp100Gly, Tyr102Ala) are shown (SDS-PAGE stained with

coomassie blue) as an example of the equimolar ratio in the intact holotoxin. **b**, Cell cycle analysis by flow cytometry of propidium iodide stained HeLa cells exposed to recombinant CDT (Methods). The calculated percentages of cells in G0/G1 and G2/M are shown.

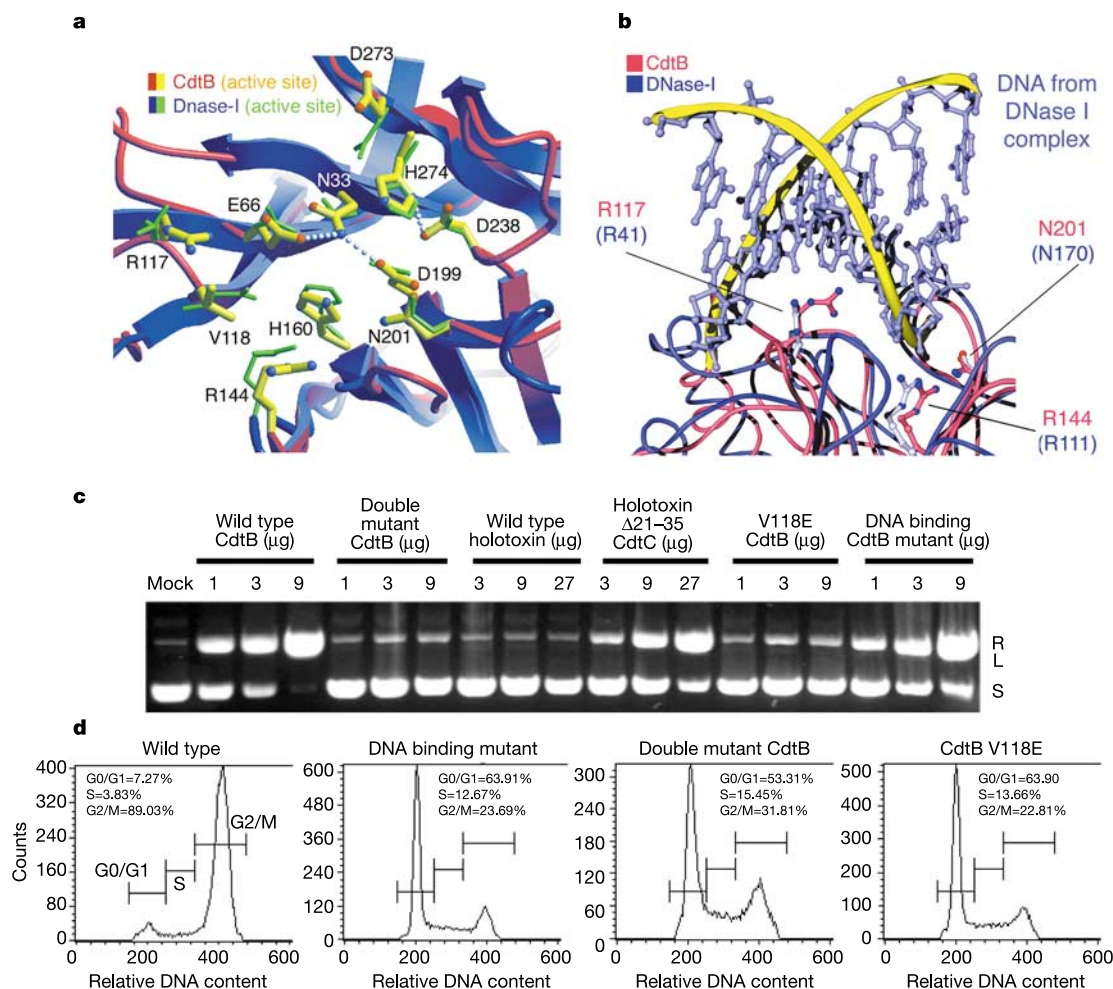


Figure 4 Comparison of CdtB with DNase I and structure-based mutants. **a**, Overall structural alignment of DNase I to CdtB generated by minimizing the C α distances. Main-chain/side-chain atoms of CdtB and DNase I are shown in red/yellow and blue/green, respectively. **b**, Comparison of the DNA contacting residues of DNase I (blue main-chain, grey side chains) with conserved residues in CdtB (magenta main-chain and side-chain). The DNA is positioned to CdtB as in Fig. 2b. **c**, *In vitro* nuclease activity of wild-type

and mutant CdtB and CDT holotoxin in plasmid-nicking assay (Methods). The three observed bands—supercoiled (S), relaxed (R) and linear (L)—are labelled. The double mutant is CdtB with His160Gln and Asp199Ser. The DNA-binding triple mutant has the mutations Arg117Ala, Arg144Ala and Asn201Ala. **d**, Cell cycle analysis following exposure to recombinant CDT with mutants in CdtB (Methods, and as in Fig. 3b).

Arg 41 (of DNase I) and Arg 117 (of CdtB) are only 2.0 Å apart and the side-chain atoms align closely (Fig. 4b), this arginine from CdtB is not located on the equivalent secondary structural element in the protein fold, but is situated on an adjacent β -strand, nearly 50 residues subsequent to the sequence position in CdtB that aligns with Arg 41 of DNase I. Other contacts between DNase I and DNA do not seem to have simple correlates in CdtB. When the three putative DNA-contacting residues are mutated, there is a complete loss of activity in cellular toxicity assays, but there are no defects in complex assembly, stability or solubility (Fig. 4d and data not shown). *In vitro* nuclease assays, against plasmid, show that the mutation of these residues diminishes activity, but, unlike corresponding mutations in DNase I, does not abolish it (Fig. 4c).

These results provide an important step towards a molecular understanding of this unique and potent toxin, as well a template for probing the function and cellular effects of CDT intoxication. In addition, the molecular details of the CdtB active site, the identification of likely substrate contacting residues, and the discovery of functionally important surfaces on the lectin domains, provide targets for therapeutic design to inactivate this toxin. □

Methods

Reconstitution and purification of the *H. ducreyi* holotoxin

H. ducreyi CdtA(18–223), CdtB(23–283) and CdtC(21–186) were cloned by PCR (polymerase chain reaction) from genomic DNA (American Type Culture Collection number 700724D) as N-terminal hexahistidine fusion proteins in *E. coli* such that the predicted N-terminal secretion signals were removed. The three subunits were expressed separately and purified under denaturing conditions (8 M urea, 10 mM Tris pH 8.0, 0.1 M Na phosphate) using nickel chelating affinity resin. Following co-refolding by dialysis at 4 °C (with total protein concentration under 0.1 mg ml⁻¹) into a native buffer consisting of 20 mM HEPES (N-[2-Hydroxyethyl]piperazine-N-[2-ethanesulphonic acid]) pH 7.5, 200 mM NaCl, 2.5 mM dithiothreitol (DTT), 5% glycerol and 2 mM ethylenediaminetetraacetic acid (EDTA), the proteins were separated from the affinity tag by site-specific proteolytic cleavage, and the holotoxin further purified by cation-exchange and gel-filtration chromatography.

Crystallization and structural determination

For crystallization, the purified preparation of the *H. ducreyi* holotoxin was concentrated by ultrafiltration to 10 mg ml⁻¹ in a buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl and 2.5 mM DTT. Crystals of the holotoxin were grown at 23 °C by vapour diffusion using hanging drops formed from mixing a 1:1 volume ratio of 10 mg ml⁻¹ complex with an equilibration buffer consisting of 15–25% polyethylene glycol monomethylether (PME-M_r 5K) 25–30% glycerol and 0.1 M imidazole, pH 6.5, supplemented with 2 mM DTT. Crystals formed in the space group P2₁2₁2₁ with cell dimensions $a = 71.7$ Å,

$b = 75.8 \text{ \AA}$ and $c = 121.8 \text{ \AA}$, with one holotoxin complex in the asymmetric unit. For cryoprotection, crystals were transferred directly into a buffer with PME (M_r 5K), concentration 5–10% higher than the conditions in which the crystals grew, with 20% glycerol, supplemented with 1 M sodium bromide, and flash-cooled immediately afterward to -160°C (total transfer/soak time between 20–30 s). Phases were determined by single-wavelength anomalous diffraction from the scattering of 26 partially ordered bromine ions introduced into the crystal from the cryo-buffer. Crystallographic methods and statistics can be found in Supplementary Information. The final model was refined to $2.0 \text{ \AA}$ and has R/R_{free} values of 18.2% and 21.4%, respectively. 90% of the residues fall into the most favourable region of the Ramachandran plot without outliers. The N-terminal 38 residues (18–56) of the crystallized construct of CdtA are not visible in the electron density, nor are residues 183–185 of CdtB (internal loop). Four residues at the N terminus and eight at the C terminus of CdtC are also not modelled, owing to disorder.

Mutagenesis

The amino-acid substitutions were introduced into *cdt* genes by *in vitro* site-directed mutagenesis using oligonucleotide primer pairs containing the appropriate base changes. The amplification of the mutant plasmid was carried out by Pfu Turbo DNA polymerase (Stratagene) in the thermal temperature cycler, and wild-type template plasmid was removed by digestion with *DpnI* before transformation. The double mutant is CdtB with His160Gln and Asp199Ser. The DNA binding mutant has the mutations Arg117Ala, Arg144Ala and Asn201Ala. The aromatic patch mutant contains Trp91Gly, Trp98Gly, Trp100Gly and Try102Ala substitutions in CdtA. The deletion mutant (CdtC Δ 21–35) was generated by PCR and cloned as N-terminal hexahistidine fusion protein into the same vector as wild-type gene. All mutants created were verified by DNA sequencing. Purification of the mutant proteins and assembly into the mutant CDT holotoxin was performed as previously described for the wild-type complex.

Cell cycle analysis

Cell cycle progression of HeLa cells treated with CDT holotoxin was analysed by flow cytometry. HeLa cells (2×10^6 cells) were treated with indicated concentrations of wild-type or mutant CDT holotoxin for 3 h at 37°C , 5% CO_2 , and were subsequently washed of toxin and maintained in culture media. 48 hours after treatment with CDT, cells were collected and fixed by addition of cold 70% ethanol during continuous vortexing, and left at 4°C for 16 h. The cells were then stained for 2 h at 23°C with propidium iodide solution ($25 \mu\text{g ml}^{-1}$ propidium iodide, 100 U ml^{-1} DNase-free RNase A, 0.04% TritonX-100 in PBS). Relative DNA content was measured by flow cytometry with FACSCalibur flow cytometer (Beckton Dickinson).

Plasmid-digestion assay

The DNase activity of CdtB was tested by a plasmid-digestion reaction as previously described^{9,30}. Wild-type and mutant CdtB proteins used in this assay were purified using the same protocol described above for the CDT holotoxin. The reaction contained 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM MgCl_2 , 5 mM CaCl_2 , supercoiled pUC19 plasmid ($2.5 \mu\text{g}$) as substrate, and indicated concentrations of wild-type or mutant CdtB (1, 3, or $9 \mu\text{g}$), or CDT holotoxin (3, 9, and $27 \mu\text{g}$). The reaction was incubated for 5 h at 37°C , and quenched by addition of 10 mM EDTA. Supercoiled, linear and relaxed plasmid forms were separated by agarose gel electrophoresis (0.8%), stained with ethidium bromide, and photographed using Gel Doc 2000 Documentation system (Bio-Rad).

Received 3 February; accepted 30 March 2004; doi:10.1038/nature02532.

- Johnson, W. M. & Lior, H. A new heat-labile cytolethal distending toxin (CLDT) produced by *Campylobacter* spp. *Microb. Pathog.* **4**, 115–126 (1988).
- Pickett, C. L. & Whitehouse, C. A. The cytolethal distending toxin family. *Trends Microbiol.* **7**, 292–297 (1999).
- De Rycke, J. & Oswald, E. Cytolethal distending toxin (CDT): a bacterial weapon to control host cell proliferation? *FEMS Microbiol. Lett.* **203**, 141–148 (2001).
- Lara-Tejero, M. & Galan, J. E. Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions. *Trends Microbiol.* **10**, 147–152 (2002).
- Lara-Tejero, M. & Galan, J. E. CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.* **69**, 4358–4365 (2001).
- Cortes-Bratti, X., Chaves-Olarte, E., Lagergard, T. & Thelestam, M. Cellular internalization of cytolethal distending toxin from *Haemophilus ducreyi*. *Infect. Immun.* **68**, 6903–6911 (2000).
- Frisan, T., Cortes-Bratti, X., Chaves-Olarte, E., Stenlerow, B. & Thelestam, M. The *Haemophilus ducreyi* cytolethal distending toxin induces DNA double-strand breaks and promotes ATM-dependent activation of RhoA. *Cell. Microbiol.* **5**, 695–707 (2003).
- Frisan, T., Cortes-Bratti, X. & Thelestam, M. Cytolethal distending toxins and activation of DNA damage-dependent checkpoint responses. *Int. J. Med. Microbiol.* **291**, 495–499 (2002).
- Elwell, C. A. & Dreyfus, L. A. DNase I homologous residues in CdtB are critical for cytolethal distending toxin-mediated cell cycle arrest. *Mol. Microbiol.* **37**, 952–963 (2000).
- Lara-Tejero, M. & Galan, J. E. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. *Science* **290**, 354–357 (2000).
- Elwell, C., Chao, K., Patel, K. & Dreyfus, L. *Escherichia coli* CdtB mediates cytolethal distending toxin cell cycle arrest. *Infect. Immun.* **69**, 3418–3422 (2001).
- Mao, X. & Di Rienzo, J. M. Functional studies of the recombinant subunits of a cytolethal distending holotoxin. *Cell. Microbiol.* **4**, 245–255 (2002).
- Cortes-Bratti, X., Frisan, T. & Thelestam, M. The cytolethal distending toxins induce DNA damage and cell cycle arrest. *Toxicol.* **39**, 1729–1736 (2001).
- Hassane, D. C., Lee, R. B., Mendenhall, M. D. & Pickett, C. L. Cytolethal distending toxin demonstrates genotoxic activity in a yeast model. *Infect. Immun.* **69**, 5752–5759 (2001).
- Cortes-Bratti, X., Karlsson, C., Lagergard, T., Thelestam, M. & Frisan, T. The *Haemophilus ducreyi* cytolethal distending toxin induces cell cycle arrest and apoptosis via the DNA damage checkpoint pathways. *J. Biol. Chem.* **276**, 5296–5302 (2001).

- Alby, F. *et al.* Study of the cytolethal distending toxin (CDT)-activated cell cycle checkpoint. Involvement of the CHK2 kinase. *FEBS Lett.* **491**, 261–265 (2001).
- Deng, K., Latimer, J. L., Lewis, D. A. & Hansen, E. J. Investigation of the interaction among the components of the cytolethal distending toxin of *Haemophilus ducreyi*. *Biochem. Biophys. Res. Commun.* **285**, 609–615 (2001).
- Lewis, D. A. *et al.* Characterization of *Haemophilus ducreyi* cdtA, cdtB, and cdtC mutants in *in vitro* and *in vivo* systems. *Infect. Immun.* **69**, 5626–5634 (2001).
- Comayras, C. *et al.* *Escherichia coli* cytolethal distending toxin blocks the HeLa cell cycle at the G2/M transition by preventing cdc2 protein kinase dephosphorylation and activation. *Infect. Immun.* **65**, 5088–5095 (1997).
- Escalas, N. *et al.* Study of the cytolethal distending toxin-induced cell cycle arrest in HeLa cells: involvement of the CDC25 phosphatase. *Exp. Cell Res.* **257**, 206–212 (2000).
- Nishikubo, S. *et al.* An N-terminal segment of the active component of the bacterial genotoxin cytolethal distending toxin B (CDTB) directs CDTB into the nucleus. *J. Biol. Chem.* **278**, 50671–50681 (2003).
- Montfort, W. *et al.* The three-dimensional structure of ricin at $2.8 \text{ \AA}$. *J. Biol. Chem.* **262**, 5398–5403 (1987).
- Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
- Lee, R. B., Hassane, D. C., Cottle, D. L. & Pickett, C. L. Interactions of *Campylobacter jejuni* cytolethal distending toxin subunits CdtA and CdtC with HeLa cells. *Infect. Immun.* **71**, 4883–4890 (2003).
- Deng, K. & Hansen, E. J. A CdtA-CdtC complex can block killing of HeLa cells by *Haemophilus ducreyi* cytolethal distending toxin. *Infect. Immun.* **71**, 6633–6640 (2003).
- Shenker, B. J. *et al.* *Actinobacillus actinomycetemcomitans* cytolethal distending toxin (Cdt): evidence that the holotoxin is composed of three subunits: CdtA, CdtB, and CdtC. *J. Immunol.* **172**, 410–417 (2004).
- Suck, D., Lahm, A. & Oefner, C. Structure refined to $2 \text{ \AA}$ of a nicked DNA octanucleotide complex with DNase I. *Nature* **332**, 464–468 (1988).
- Weston, S. A., Lahm, A. & Suck, D. X-ray structure of the DNase I-d(GGTATACC)₂ complex at $2.3 \text{ \AA}$ resolution. *J. Mol. Biol.* **226**, 1237–1256 (1992).
- Jones, S. J., Worrall, A. F. & Connolly, B. A. Site-directed mutagenesis of the catalytic residues of bovine pancreatic deoxyribonuclease I. *J. Mol. Biol.* **264**, 1154–1163 (1996).
- Pan, C. Q., Ulmer, J. S., Herzka, A. & Lazarus, R. A. Mutational analysis of human DNase I at the DNA binding interface: implications for DNA recognition, catalysis, and metal ion dependence. *Protein Sci.* **7**, 628–636 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Mueller and T. Radhakannan for access to and assistance with crystallographic equipment, and S. Mazel for access to a flow cytometer. This work was funded by research funds to C.E.S. from the Rockefeller University.

Authors' contributions D.N.—cloning of wild-type and mutant CDT holotoxin and CdtB, protein purification, activity assays, and crystallography, and Y.H.—mutant CdtA cloning and purification of mutant CdtA containing holotoxin.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.E.S. (stebbins@rockefeller.edu). Coordinates of the structure have been deposited in the Protein Data Bank under the accession number 1SR4.

Human meiotic recombinase Dmc1 promotes ATP-dependent homologous DNA strand exchange

Michael G. Sehorn^{1*}, Stefan Sigurdsson^{2*}, Wendy Bussen²,
Vinzenn M. Unger¹ & Patrick Sung¹

¹Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA

²Department of Molecular Medicine, Institute of Biotechnology, University of Texas Health Science Center at San Antonio, 15355 Lambda Drive, San Antonio, Texas 78245, USA

* These authors contributed equally to this work

Homologous recombination is crucial for the repair of DNA breaks and maintenance of genome stability^{1–3}. In *Escherichia coli*, homologous recombination is dependent on the RecA protein. In the presence of ATP, RecA mediates the homologous DNA pairing and strand exchange reaction that links recombin-

ing DNA molecules. DNA joint formation is initiated through the nucleation of RecA onto single-stranded DNA (ssDNA) to form helical nucleoprotein filaments^{4–6}. Two RecA-like recombinases, Rad51 and Dmc1, exist in eukaryotes¹. Whereas Rad51 is needed for both mitotic and meiotic recombination events, the function of Dmc1 is restricted to meiosis^{3,7}. Here we examine human Dmc1 protein (hDmc1) for the ability to promote DNA strand exchange, and show that hDmc1 mediates strand exchange between paired DNA substrates over at least several thousand base pairs. DNA strand exchange requires ATP and is strongly dependent on the heterotrimeric ssDNA-binding molecule replication factor A (RPA). We present evidence that hDmc1-mediated DNA recombination initiates through the nucleation of hDmc1 onto ssDNA to form a helical nucleoprotein filament. The DNA strand exchange activity of hDmc1 is probably indispensable for repair of DNA double-strand breaks during meiosis and for maintaining the ploidy of meiotic chromosomes.

The *RAD51* gene is required for mitotic and meiotic recombination^{3,8}. Like RecA, Rad51 forms helical nucleoprotein filaments and mediates homologous DNA pairing and strand exchange in an ATP-dependent manner^{2,4}. Mutations in *DMC1* (disrupted meiosis cDNA 1) cause no discernible mitotic phenotype but lead to a constellation of meiotic abnormalities, primarily due to a failure to pair homologous chromosomes properly^{7,9–11}. Surprisingly, even

though Dmc1 protein has a weak homologous DNA pairing activity, it does not seem to catalyse DNA strand exchange^{1,12–14}. The apparent lack of DNA strand exchange activity in Dmc1 is counter-intuitive, because genetic studies in yeast have shown that *DMC1* deletion suppresses meiotic recombination, and Dmc1 protein seems to form recombinants in the absence of Rad51 (refs 7, 15). Additionally, Dmc1 is thought to bind DNA as stacked protein rings and to mediate homologous DNA pairing by means of this nucleoprotein intermediate^{13,16}.

We cloned hDmc1 cDNA and introduced it into a baculovirus vector that adds a 5' six-histidine tag. A recombinant His₆-hDmc1 baculovirus was made and used to infect High-Five insect cells. We purified hDmc1 by three chromatographic fractionation steps and affinity chromatography on Ni²⁺-nitrilotriacetate agarose (Fig. 1a); the identity of the purified protein was verified by matrix-assisted laser desorption ionization–time-of-flight MS analysis. Three different hDmc1 preparations gave the same results in all the analyses.

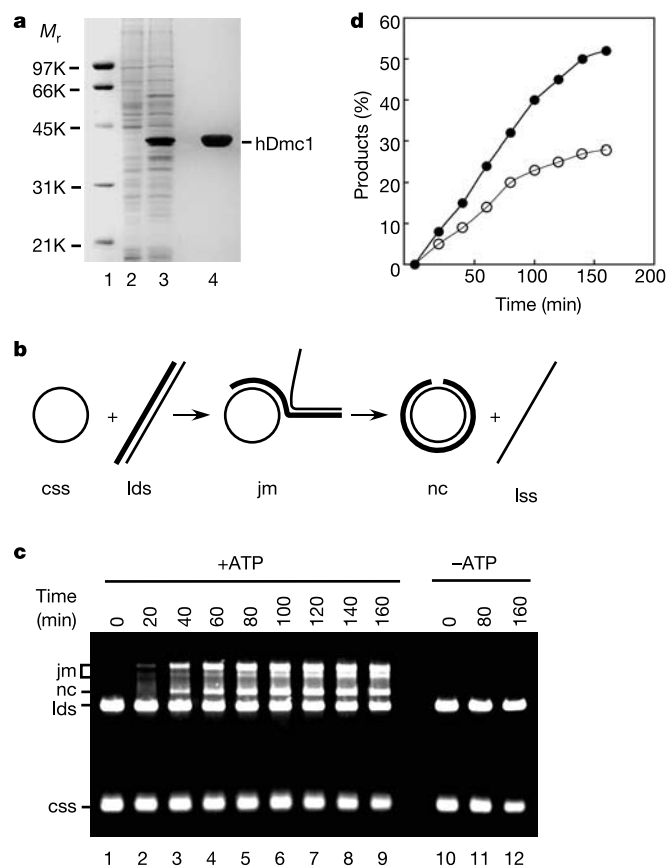


Figure 1 hDmc1 catalyses ATP-dependent DNA strand exchange. **a**, Extracts from cells without (lane 2) or with (lane 3) hDmc1, and purified hDmc1 (4 μ g in lane 4), were analysed by SDS–polyacrylamide-gel electrophoresis on 10% polyacrylamide gels. M_r , relative molecular mass. **b**, Reaction diagram: pairing of (+) strand (css) with linear duplex (lds) yields a joint molecule (jm), and DNA strand exchange generates a nicked circular duplex (nc) and linear ssDNA (lss). **c**, DNA strand exchange as a function of reaction time (lanes 1–9). ATP is needed for product formation (lanes 10–12). **d**, Time course of DNA strand exchange. Open circles, nicked circular duplex; filled circles, joint molecules and nicked circular duplex combined.

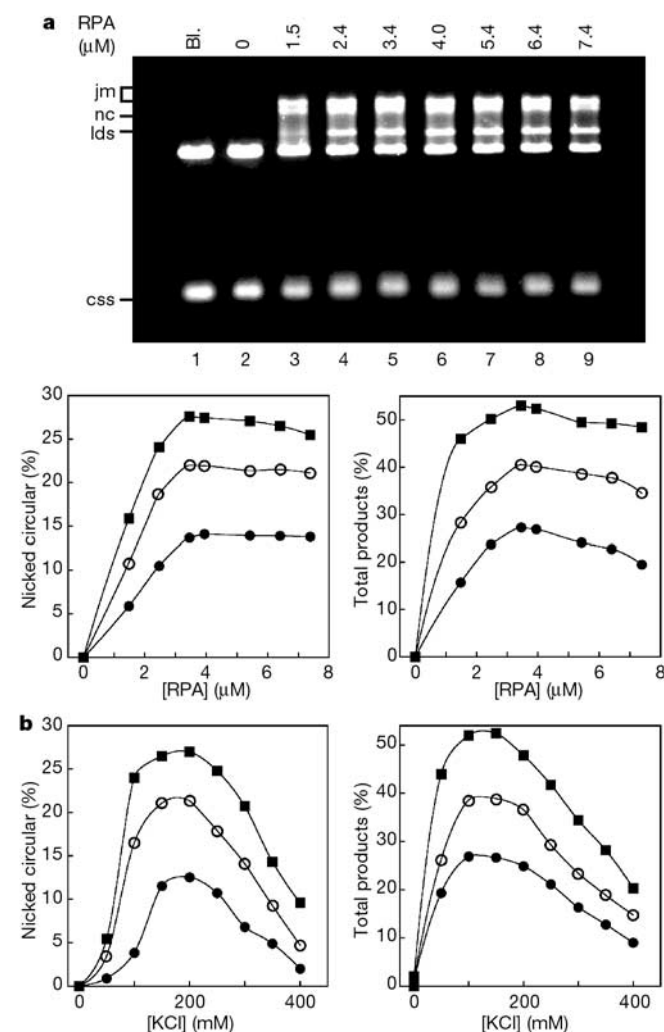


Figure 2 Dependence of hDmc1-mediated DNA strand exchange on hRPA and salt. **a**, Dependence of hDmc1-mediated DNA strand exchange on hRPA. The agarose gel containing samples from the 3-h time point is shown in the upper panel; labels are as described in Fig. 1b. Lower left, nicked circular duplex; lower right, nicked circular duplex and joint molecules combined. **b**, Dependence of hDmc1/hRPA-mediated DNA strand exchange on KCl. Left, nicked circular duplex; right, nicked circular duplex and joint molecules combined. Symbols: filled circles, 60 min of incubation; open circles, 120 min of incubation; squares, 180 min of incubation.

We examined whether hDmc1 can mediate DNA strand exchange with plasmid-length ϕ X174 ssDNA and double-stranded DNA (dsDNA)²⁻⁴. Pairing yields a DNA joint molecule, and DNA strand exchange over the length (5.4 kilobases) of the synapsed substrates produces nicked circular duplex and linear ssDNA²⁻⁴ (Fig. 1b). We included hRPA in the reaction, because this ssDNA-binding factor enhances the Rad51 recombinase activity²⁻⁴. We first verified that the hDmc1 and hRPA used were devoid of nuclease contamination by incubating them with the appropriate test substrates (data not shown). For DNA strand exchange, we preincubated hDmc1 with

ssDNA in the presence of ATP, and then incorporated hRPA and dsDNA. A substantial level of nicked circular duplex was seen. Specifically, at 80 min, about 20% of the input duplex was converted into nicked circular duplex (Fig. 1c). The use of duplex substrates labelled with ³²P at either the 3' or 5' end confirmed the displacement of linear ssDNA and provided assurance that hDmc1 catalyses genuine DNA strand exchange (Supplementary Fig. 1). Maximal DNA strand exchange occurs within the range of two nucleotides to one nucleotide per hDmc1 monomer (data not shown).

No product was seen when ATP was omitted (Fig. 1c) or when it was replaced with ADP or one of the slowly hydrolysable ATP analogues ATP- γ S, AMP-PNP or AMP-PCP (data not shown). Thus, ATP hydrolysis seems to be indispensable for the hDmc1-mediated recombination reaction. As shown in Fig. 2a, the omission of hRPA completely abolished the formation of nicked circular duplex and greatly decreased the amount of DNA joint molecules. The optimal level of hRPA for the hDmc1-mediated reaction was between 10 and 15 nucleotides per RPA monomer (Fig. 2a). As expected, no product was seen with heterologous DNA species (for example, ϕ X (+) strand and linear pBluescript duplex), confirming the requirement for DNA homology (data not shown). Interestingly, physiological ionic strength (100–200 mM KCl) is essential for revealing the recombinase activity of hDmc1 (Fig. 2b). The recombination reaction is also highly sensitive to pH, with optimal activity being observed between pH 7.3 and 7.8 (data not shown).

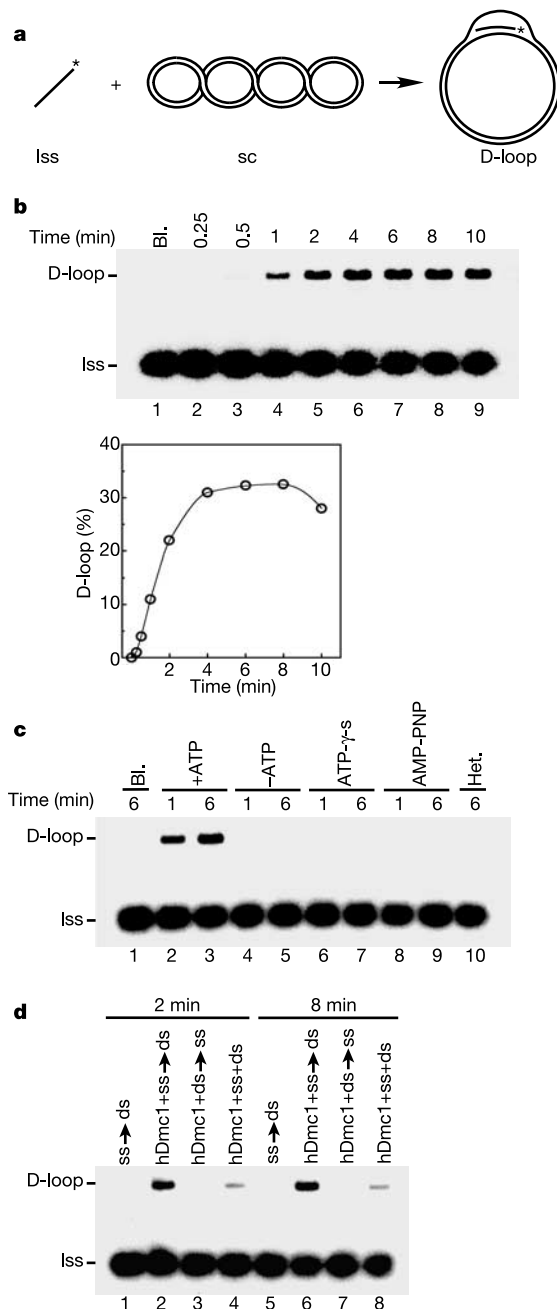


Figure 3 hDmc1 mediates robust D-loop formation. **a**, Reaction diagram: pairing of OL2 (lss) with pBluescript replicative form I DNA (sc) yields D-loop. **b**, D-loop formation as a function of time. BL, reaction blank without hDmc1. **c**, The dependence of D-loop formation on ATP and DNA homology. The heterologous DNA (ϕ X174 replicative form I DNA, designated as Het in lane 10) was used with hDmc1 and ATP. BL, reaction blank with the homologous DNA substrates but no hDmc1. **d**, The level of D-loop was examined as a function of the order of addition of reaction components.

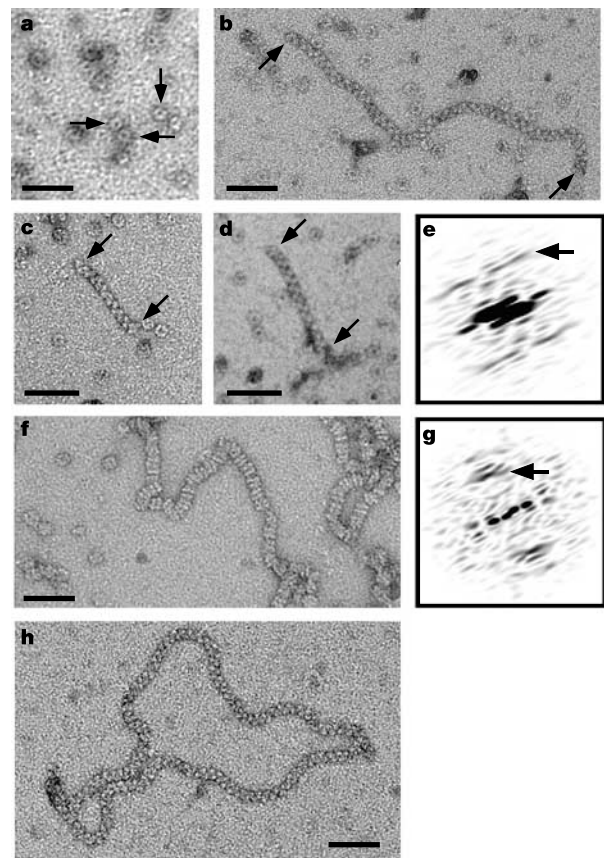


Figure 4 Analysis of hDmc1-ssDNA complexes by electron microscopy. **a**, hDmc1 protein rings (several marked by arrows). **b–d**, hDmc1-ssDNA filaments formed with ATP; the filament ends are marked by arrows. **e**, Fourier transform of the filament in **d**. A pair of layer lines (marked by the arrow) was observed 1/82 Å from the equator, indicating helical symmetry. **f**, Without ATP, only stacked rings of hDmc1 are formed on ssDNA. **g**, Fourier transform of the stacked rings shows one strong layer line (marked by the arrow) crossing the meridian. **h**, A helical hRad51-ssDNA filament formed with ATP. Scale bar, 50 nm.

We also examined the ability of hDmc1 to catalyse DNA pairing and strand invasion in the D-loop assay, which monitors the assimilation of a radiolabelled oligonucleotide into a homologous duplex²⁻⁴ (Fig. 3a). The hDmc1-mediated D-loop reaction is robust (Fig. 3b) and requires ATP, which cannot be replaced by ATP- γ S or AMP-PNP (Fig. 3c). Importantly, preincubation of hDmc1 with ssDNA resulted in a robust reaction, whereas less product was seen when the duplex substrate was preincubated with hDmc1, or when the recombinase was mixed with both DNA substrates simultaneously (Fig. 3d). These results implicate a hDmc1-ssDNA complex as the active nucleoprotein species in recombination, a conclusion that was independently verified by experiments on the order of addition with the use of the ϕ X174 DNA substrates (data not shown). As shown in Supplementary Fig. 2, the ability of hDmc1 to form D-loop is enhanced by the Rad54B protein, which is probably the orthologue of the yeast Rdh54/Tid1 protein known to be involved in meiotic recombination with Dmc1 (ref. 3).

The functional form of hDmc1 is thought to comprise stacked protein rings on DNA^{13,16}. We used electron microscopy to examine the nature of the hDmc1-ssDNA nucleoprotein complex, adhering to the buffer conditions for DNA strand exchange. Consistent with previous studies^{13,16} was our observation that only hDmc1 protein rings were seen in the absence of DNA (Fig. 4a). hDmc1 protein filaments were formed upon the addition of ssDNA; three examples of these filaments are shown in Fig. 4b-d. The hDmc1-ssDNA nucleoprotein filaments had the striated appearance characteristic of helical nucleoprotein filaments made by Rad51 (Fig. 4h) and RecA^{1,4,6}. A Fourier transform further confirmed that the hDmc1-ssDNA nucleoprotein filaments were helical (Fig. 4e). Although abundant, the hDmc1-ssDNA helical nucleoprotein filaments were shorter than those assembled with hRad51, with a mean length of 98 ± 43 nm ($n = 152$), although longer filaments were seen occasionally (see Fig. 4b, for example). Importantly, from several independent spreads, we found no evidence of stacked hDmc1 rings when ATP was present. However, when ATP was omitted there was no sign of hDmc1 helical filaments; instead, linear arrays of stacked hDmc1 rings on ssDNA were seen (Fig. 4f); a Fourier transform helped to establish the non-helical nature of these nucleoprotein complexes (Fig. 4g).

In recombination reactions, ssDNA is used to initiate genetic exchange with a homologous duplex²⁻⁵. RecA and Rad51 function in recombination by assembling into a helical protein filament on ssDNA often referred to as the presynaptic filament^{2,4}. The presynaptic filament contains a binding site for duplex DNA, and this ability of the presynaptic filament to hold three DNA strands in close proximity underlies the basis for seeking out DNA homology and forming DNA joints^{2,4}. Here we show that hDmc1 also forms a helical nucleoprotein filament on ssDNA and mediates extensive DNA strand exchange. In experiments that used fluorescence resonance energy transfer, hDmc1 was seen to form a three-stranded synaptic intermediate¹⁷. It is likely that the hDmc1-ssDNA nucleoprotein filament represents the critical catalytic intermediate for this recombinase. That the hDmc1 filaments in our electron microscopic analysis did not extend over the entire length of the DNA might reflect an inherent instability of these filaments. We find that a relatively high concentration of salt and a narrow pH range are required to reveal the DNA strand exchange potential of hDmc1. It remains to be seen whether these reaction parameters help to modulate the affinity of the hDmc1 presynaptic filament for duplex DNA or affect another step in the reaction path^{2,18}.

Our results show that hRPA is critical for DNA strand exchange efficiency. hRPA could promote hDmc1-mediated DNA strand exchange by minimizing secondary structure in the ssDNA²⁻⁴ and sequestering the displaced strand that results from DNA strand exchange^{19,20}. In addition, we have found that hRad54B protein enhances the ability of hDmc1 to form D-loop. It is very likely that the hDmc1 recombinase activity is modulated by other protein

factors that function with Dmc1 in meiosis²¹; in this regard, the biochemical systems devised here will be valuable for defining the role of these factors in hDmc1-mediated recombination reactions. $\square$

Methods

Purification of hDmc1 and hRPA

High-Five insect cells were infected with His₆-hDmc1 baculovirus at a multiplicity of infection of 10 and were harvested after 48 h. Extract from 200 ml of insect cell culture was subjected to chromatographic fractionation in Q Sepharose, Mono Q and Mono S, and also to affinity purification with Ni²⁺-nitrilotriacetate agarose (Qiagen) to purify hDmc1 to near homogeneity. A total of 5 mg protein was obtained, which was concentrated in a Centricon-30 microconcentrator to 15 mg ml⁻¹ and stored in small aliquots at -70 °C. Human RPA was purified to near homogeneity from *E. coli*²² as described¹⁸.

DNA substrates

The ϕ X174 circular ssDNA and replicative form I DNA (about 95% supercoiled) were purchased from New England Biolabs and Life Technologies, respectively. The replicative form I DNA was linearized with *Apa*L1. OL2 is a 90-mer oligonucleotide complementary to pBluescript SK DNA and was ³²P-labelled at the 5' end as described²³. DNA substrates were stored in TE buffer (10 mM Tris-HCl pH 7.2, 0.5 mM EDTA).

DNA strand exchange reaction

The cited reactant concentrations corresponded to the final values and the incubation steps were performed at 37 °C. The time course experiment shown in Fig. 1 had a final volume of 50 μ l and was performed in 35 mM Tris-HCl pH 7.6, 1 mM dithiothreitol, 2 mM ATP, 2.5 mM MgCl₂, with an ATP-regenerating system consisting of 8 mM creatine phosphate and 30 μ g ml⁻¹ creatine kinase. To assemble the reaction, hDmc1 (30 μ M) was incubated with ϕ X174 (+) strand (30 μ M nucleotides) for 5 min. At this time, hRPA (2 μ M) and KCl (200 mM) were added, and after a 5-min incubation, linear ϕ X dsDNA (22 μ M nucleotides) was incorporated to complete the reaction. At the indicated times, a 5 μ l aliquot of the reaction mixture was withdrawn, deproteinized and subjected to analysis in agarose gels, as described²⁴. In other experiments the volume of the reaction was scaled down appropriately. The experiments in Fig. 2 used the same hDmc1 and DNA concentrations as above.

D-loop reaction

The reaction buffer was 20 mM Tris-HCl pH 7.3, 100 μ g ml⁻¹ BSA, 5 mM MgCl₂, 2 mM ATP, containing the ATP-generating system described above. For the experiment in Fig. 3 (37.5 μ l final volume), hDmc1 (6.5 μ M) was incubated with 5'-labelled OL2 (2.5 μ M nucleotides) for 3 min at 37 °C. The reaction was completed with the addition of pBluescript SK replicative form I DNA (190 μ M nucleotides). The reaction mixture was incubated at 30 °C, and 3.3 μ l aliquots were withdrawn at the indicated times, deproteinized, and run in 1% agarose gels in Tris/acetate/EDTA buffer as described²³. The gels were dried and D-loop levels were quantified by phosphorimaging analysis. The reactions in which ATP was omitted or replaced by a non-hydrolysable analogue were scaled down to 12.5 μ l final volume.

Electron microscopy and image analysis

Reaction mixtures were assembled in the buffer used for DNA strand exchange. The reactions contained hDmc1 (30 μ M) or Rad51 (30 μ M) and ssDNA (120 μ M nucleotides) in the absence or presence of 2 mM ATP and were incubated for 1 h at 37 °C. After dilution 1:40 with the reaction buffer, 3 μ l of the diluted mixture was applied to 400-mesh grids coated with fresh carbon film that had been glow-discharged in air. After staining for 15 s with 2% uranyl acetate, samples were examined in a Tecnai 12 transmission electron microscope equipped with a LaB6 filament and operated at 120 keV. Digital images were captured with a GATAN 794, 1,024 $\times$ 1,024-pixel charge-coupled device camera at a nominal magnification of $\times 26,000$. For further analysis, images were converted to Medical Research Council (MRC) format, and ordered segments of the nucleoprotein complexes were identified with the fast Fourier transform option in the image display program XIMDISP²⁵. Diffraction patterns were calculated after boxing short, straight segments of the nucleoprotein complexes with the program BOXIMAGE²⁵.

Received 27 March; accepted 14 April 2004; doi:10.1038/nature02563.

1. Masson, J. Y. & West, S. C. The Rad51 and Dmc1 recombinases: a non-identical twin relationship. *Trends Biochem. Sci.* **26**, 131-136 (2001).
2. Sung, P., Krejci, L., Van Komen, S. & Sehorn, M. G. Rad51 recombinase and recombination mediators. *J. Biol. Chem.* **278**, 42729-42732 (2003).
3. Symington, L. S. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiol. Mol. Biol. Rev.* **66**, 630-670 (2002).
4. Bianco, P. R., Tracy, R. B. & Kowalczykowski, S. C. DNA strand exchange proteins: a biochemical and physical comparison. *Front. Biosci.* **3**, D570-D603 (1998).
5. Roca, A. I. & Cox, M. M. RecA protein: structure, function, and role in recombinational DNA repair. *Prog. Nucleic Acid Res. Mol. Biol.* **56**, 129-223 (1997).
6. Yu, X., Jacobs, S. A., West, S. C., Ogawa, T. & Egelman, E. H. Domain structure and dynamics in the helical filaments formed by RecA and Rad51 on DNA. *Proc. Natl Acad. Sci. USA* **98**, 8419-8424 (2001).
7. Bishop, D. K., Park, D., Xu, L. & Kleckner, N. DMC1: a meiosis-specific yeast homolog of *E. coli* recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69**, 439-456 (1992).
8. Shinohara, A. et al. Cloning of human, mouse and fission yeast recombination genes homologous to RAD51 and recA. *Nature Genet.* **4**, 239-243 (1993).

9. Lydall, D., Nikolsky, Y., Bishop, D. K. & Weinert, T. A meiotic recombination checkpoint controlled by mitotic checkpoint genes. *Nature* **383**, 840–843 (1996).
10. Yoshida, K. *et al.* The mouse RecA-like gene Dmc1 is required for homologous chromosome synapsis during meiosis. *Mol. Cell* **1**, 707–718 (1998).
11. Pittman, D. L. *et al.* Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for Dmc1, a germline-specific RecA homolog. *Mol. Cell* **1**, 697–705 (1998).
12. Li, Z., Golub, E. I., Gupta, R. & Radding, C. M. Recombination activities of HsDmc1 protein, the meiotic human homolog of RecA protein. *Proc. Natl Acad. Sci. USA* **94**, 11221–11226 (1997).
13. Masson, J. Y. *et al.* The meiosis-specific recombinase hDmc1 forms ring structures and interacts with hRad51. *EMBO J.* **18**, 6552–6560 (1999).
14. Hong, E. L., Shinohara, A. & Bishop, D. K. *Saccharomyces cerevisiae* Dmc1 protein promotes renaturation of single-strand DNA (ssDNA) and assimilation of ssDNA into homologous supercoiled duplex DNA. *J. Biol. Chem.* **276**, 41906–41912 (2001).
15. Shinohara, A., Ogawa, H. & Ogawa, T. Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. *Cell* **69**, 457–470 (1992).
16. Passy, S. I. *et al.* Human Dmc1 protein binds DNA as an octameric ring. *Proc. Natl Acad. Sci. USA* **96**, 10684–10688 (1999).
17. Gupta, R. C., Golub, E., Bi, B. & Radding, C. M. The synaptic activity of HsDmc1, a human recombination protein specific to meiosis. *Proc. Natl Acad. Sci. USA* **98**, 8433–8439 (2001).
18. Sigurdsson, S., Trujillo, K., Song, B., Stratton, S. & Sung, P. Basis for avid homologous DNA strand exchange by human Rad51 and RPA. *J. Biol. Chem.* **276**, 8798–8806 (2001).
19. Van Komen, S., Petukhova, G., Sigurdsson, S. & Sung, P. Functional crosstalk among Rad51, Rad54, and RPA in heteroduplex DNA joint formation. *J. Biol. Chem.* **277**, 43578–43587 (2002).
20. Eggler, A. L., Inman, R. B. & Cox, M. M. The Rad51-dependent pairing of long DNA substrates is stabilized by replication protein A. *J. Biol. Chem.* **277**, 39280–39288 (2002).
21. Zickler, D. & Kleckner, N. Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* **33**, 603–754 (1999).
22. Henriksen, L. A., Umbricht, C. B. & Wold, M. S. Recombinant replication protein A: expression, complex formation, and functional characterization. *J. Biol. Chem.* **269**, 11121–11132 (1994).
23. Sigurdsson, S., Van Komen, S., Petukhova, G. & Sung, P. Homologous DNA pairing by human recombination factors Rad51 and Rad54. *J. Biol. Chem.* **277**, 42790–42794 (2002).
24. Sung, P. & Robberson, D. L. DNA strand exchange mediated by a RAD51-ssDNA nucleoprotein filament with polarity opposite to that of RecA. *Cell* **82**, 453–461 (1995).
25. Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 9–16 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

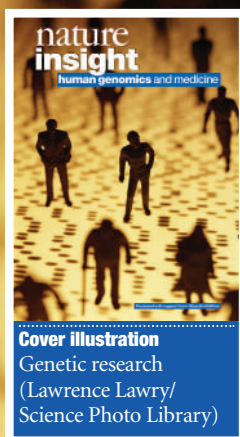
Acknowledgements We thank T. Habu for helping with cDNA cloning and baculovirus construction. This work was supported by grants from the US National Institutes of Health (P.S. and V.M.U.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.S. (patrick.sung@yale.edu).

nature insight

human genomics and medicine



Last year the official completion of the human genome sequence was announced, capping years of hard work. Now that the flashbulbs have dimmed, scientists are taking a hard look at the results. Benefits of the sequence were prophesied to include 'magic bullet' therapeutics, individualized medicine, and the prediction of disease long before symptoms surface. But to realize these breakthroughs, we must fashion the four-letter code we all share into tools physicians can use, and ensure that these tools are readily available. Although one graduate student can now make huge strides with access to the Internet and basic molecular biology equipment, true success may demand nothing short of entirely new methods of clinical study and reorganization of existing academic structures.

As the articles in this *Nature Insight* demonstrate, the intersection of genomics and clinical practice is a busy and fascinating place. The road to treatments based on genomic information has not been smooth, but clinical trials are underway for a number of new therapies. New methods of mutation screening are emerging, both for the genome and for the 'epigenome' layered onto it. An understanding of the many mutations that underlie complex diseases and adverse drug reactions is now in our sights, aided by a large international effort to define the differences between our individual genomes. And the availability of genome sequence from our close and far evolutionary relatives is helping us to decipher the signals of genes and regulatory elements from the noise of background DNA, which is not yet fully understood.

The genome community's strong tradition is a rich debate of ideas; we have captured some of them in this collection and hope to stimulate more. We are pleased to acknowledge the financial support of GlaxoSmithKline in producing this *Insight*. As always, *Nature* carries sole responsibility for all editorial content and peer review. You'll find our comprehensive human genome resource at <http://www.nature.com/nature/focus/humangenome>, including the papers published in this very issue on the sequences of human chromosomes 9 and 10.

Chris Gunter Senior Editor

Editor, *Nature*: Philip Campbell

Publisher: Peter Collins

Insights Editor: Lesley Anson

Editorial Assistant: Timothy Gibbs

Production Editors: Maria Hodges

Rebecca Craven

Art Editor: Martin Harrison

Layouts: Alisdair Macdonald

Diagrams:

Charlotte Cartwright

Debbie Maizels

Nik Spencer

Radha Clelland

Production:

Sue Gray

Marketing:

Claire Aspinall

Sponsorship:

Mark Greene

Sarah Greaves

overview:

440 Genomes for medicine

D. R. Bentley

review articles:

446 Mapping complex disease loci in whole-genome association studies

C. S. Carlson, M. A. Eberle, L. Kruglyak & D. A. Nickerson

453 Predicting disease using genomics

J. Bell

457 Epigenetics in human disease and prospects for epigenetic therapy

G. Egger, G. Liang, A. Aparicio & P. A. Jones

464 Moving towards individualized medicine with pharmacogenomics

W. E. Evans & M. V. Relling

469 Oncogenomics and the development of new cancer therapies

R. L. Strausberg, A. J. G. Simpson, L. J. Old & G. J. Riggins

commentary:

475 The case for a US prospective cohort study of genes and environment

F. S. Collins

commentary:

478 Organizational challenges in clinical genomic research

J. S. Altshuler & D. Altshuler

Genomes for medicine

David R. Bentley

The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK (e-mail: drb@sanger.ac.uk)

We have the human genome sequence. It is freely available, accurate and nearly complete. But is the genome ready for medicine? The new resource is already changing genetic research strategies to find information of medical value. Now we need high-quality annotation of all the functionally important sequences and the variations within them that contribute to health and disease. To achieve this, we need more genome sequences, systematic experimental analyses, and extensive information on human phenotypes. Flexible and user-friendly access to well-annotated genomes will create an environment for innovation, and the potential for unlimited use of sequencing in biomedical research and practice.

The DNA sequence of *Homo sapiens* is freely available and in the public domain in order to encourage research and development and to maximize its benefit to society. The commitment to this goal was made by all participants of the Human Genome Project in the Bermuda Statement in February 1996 (refs 1, 2). The successful outcome represents a remarkable achievement brought about by international cooperation, scientific excellence and altruism. It has been accompanied by striking technological progress — in DNA manipulation, automation of biological processes and computational methods for handling very large and complex sets of data. The key to the success of this endeavour lies in the fundamental nature of the information itself. No other project could produce a single data set that encompasses the genetic basis of being human. In a tidal wave of optimism, similar principles have been extended to other genomes², with the result that we now have an unrivalled foundation for biological research in the future — all the genetic information used to make humans^{3–5}, rodents^{6,7}, flies⁸, worms⁹, plants¹⁰, bacteria^{11–13} and many more. This is a revolution in knowledge that promises to change our way of thinking.

The foresight of sequencing a complete genome (as opposed to cataloguing all the available messenger RNAs) came in recognizing that a genome is more than a bundle of genes: the organization of genes in the context of surrounding information in the rest of the DNA might be important. It was worth characterizing everything precisely because we did not understand it. Some thought that it was premature to write noncoding DNA off as 'junk', and so it has proved. An enormous amount of functionally important information is now being found in addition to the protein-coding sequences. The human genome sequence contains non-coding RNA genes, regulatory sequences and structural motifs; it maintains short-range and long-range spatial organization of sequences; and it contains important evolutionary information. The genome sequence also provides a record of the natural chromosomal organization of genetic material. Given that all this information is in the sequence, it is important not to miss anything. Only by going systematically along each chromosome from end to end could every piece of information be captured with certainty. The realization that this was possible for any organism gave rise to the era of large-scale genome sequencing, which started in earnest only ten years ago.

The essential properties of the human genome sequence — accuracy and completeness — reflect intrinsic features of the molecular structure of DNA¹⁴. Physically, the genome is a continuous thread on each chromosome, with the identity and order of each base determined unequivocally by the

atomic structure of the double helix. The information stored within it is digital and can therefore be decoded unambiguously. The current version of the genome sequence ('build 34', <http://ncbi.nlm.nih.gov>) comprises 2.84 gigabases (Gb); it is more than 99.995% accurate and covers 99% of the euchromatin⁵, and therefore constitutes a high-quality reference for future work. A few parts of the thread remain intractable to current techniques and are targets for further research. They comprise the highly repetitive sequence of the heterochromatin plus the few remaining gaps in the euchromatin, many of which are believed to be rich in G+C content or repetitive sequence.

Is the genome ready for medicine?

Disease is a malfunction of the human body. It is caused by one or more internal changes, usually in combination with external factors. The human genome sequence will have a profound impact on our understanding of diseases with a genetic component. Ultimately it should be possible to examine an individual's genetic make-up at any position in the sequence, deduce a functional consequence, and make a well-informed choice of medical action. This level of understanding requires a detailed knowledge of the information in the human genome. How will this be achieved?

With a near-complete sequence in hand, we can draw together all the available human genetic information for the first time. About 22,000 protein-coding genes have been identified so far (http://www.ensembl.org/Homo_sapiens), and the sequence continues to be searched methodically for clues to any previously undiscovered members of a particular protein family. For example, we can list every protein kinase or transcription factor, or zoom in on a chromosomal region of interest and pick off every gene. Complete gene lists will enhance any experiment to search for medically important genes. With every human gene on an expression microarray, we can investigate the full extent of transcriptional changes that accompany tumorigenesis. We can use DNA arrays or comparative genome hybridization to search the entire genome of an individual for the germline variants, such as insertions, deletions, amplifications or translocations, that might be associated with phenotypes such as mental disorder or congenital abnormality. The same approach can be used to determine the combination of germline and somatic variants and somatic changes that might lead to cancer (see review in this issue by Strausberg *et al.*, page 469).

We believe that the genome contains over ten million 'common' polymorphic sites (sites where the minor allele is present in at least 1% in the human population) and an almost unlimited number of rarer variants¹⁵. So far (as at

February 2004), more than seven million variants, most of which are single-nucleotide polymorphisms (SNPs), have been catalogued in the public databases and mapped on the genome sequence^{16–18} (<http://www.ncbi.nlm.nih.gov/SNP>). We can examine each gene for variants that alter protein-coding sequence or splice sites, and test them directly to determine their functional significance. We can select polymorphisms for use as genetic markers, download the flanking sequence, develop experimental assays to determine the genotype of individual DNA samples, and search for associations with disease.

More than 1,400 human genes have been correlated directly with disease (data from Online Mendelian Inheritance in Man at <http://www.ncbi.nlm.nih.gov/Entrez>). In general these are single-gene disorders. Almost every study has pinpointed a causative mutation and implicated a specific protein that is altered or absent in the disease phenotype. Such discoveries can lead to the provision of a precise predictive test, particularly for monogenic disorders such as cystic fibrosis or Huntington's disease, and also stimulate targeted research towards an effective cure by correction or replacement of the defective protein. The discovery that a chromosome translocation creates a new gene structure (the *abl-bcr* gene) in chronic myeloid leukaemia led to the development of the drug imatinib (Gleevec), which binds specifically to the ABL-BCR protein and can alleviate the leukaemia in patients for whom other treatments have failed¹⁹. Effective cures are not necessarily guaranteed, of course: the defective protein might be inaccessible or impossible to replace, the defect might be lethal too early in life for administration of treatment, or there might be unexpected complication with the therapy. Despite early success in the treatment of severe combined immunodeficiency by replacement of the defective adenosine deaminase gene²⁰, it is proving a challenge to continue translation into clinical practice²¹, illustrating that further research is needed to relate genomic data to patients.

What about the more complex diseases, such as diabetes, heart disease, cancer or schizophrenia? We could build up the picture step by step, discovering all the genetic variants and environmental factors that contribute to the disease, and then work out all the permutations that are significant — a monumental task. However, we might not need to do this. If we identify the key pathways involved, we might be able to pinpoint the most effective points for intervention on the basis of biochemical knowledge, and avoid characterizing many of the contributing risk factors. The new targets then become the focus for the development of new drugs, leading to effective treatments.

To find a pathway, it is necessary to identify at least one of its components — for example, an enzyme in a metabolic pathway, a receptor or transducer in a signalling pathway, or a polymerase in a DNA repair pathway. Genetic approaches to this problem can benefit enormously from the human genome sequence. We can choose a gene, pick polymorphic markers from the sequence, test them for association with the disease, and then search the region for causative variants. If the disease has a familial mode of inheritance, it might be possible to use linkage analysis. If not, as has been true in most studies of common disease so far, the alternative is to use a population-based association study and look for an imbalance in allele frequencies of a marker in a group of unrelated cases compared with a matched control group²². For example, an association study demonstrated the protective effect of the 32-base deletion ($\Delta 32$) in the cytokine receptor 5 (CKR5) gene against HIV-1 infection or AIDS progression²³. In an extensive gene-based survey, two SNPs in the LTA-3 gene showed significant association with myocardial infarction²⁴. The genome is ready for us to make a start, aided by continuing efforts to complete the gene annotation, to expand the collections of variants in the public databases and to characterize the patterns of common variation in human population groups¹⁸. These developments are expected to help us dissect out the genetic basis of complex traits (see reviews in this issue by Carlson *et al.* (page 446) and Bell (page 453)), including variable drug response (see review in this issue by Evans and Relling, page 464).

Where we fall short at present is in our understanding of physiological function or how sets of molecules work together, and in our ability to infer this from the accumulated information linked to the genome. We lack knowledge of the functional sequences outside genes, as well as a detailed understanding of when and where genes are expressed, and in response to what signals. We are unaware of the biochemical functions of most proteins, and lack knowledge of most of the interactions between cellular components. If we were to acquire these information sets we would be able to go much further using the genome. Finding a genetic target would then allow us to jump immediately into a completely characterized biochemical pathway, to understand the functional processes that are disrupted by a particular mutation, and to develop measures to discern the influence of non-genetic factors on these processes.

A few examples illustrate the roadblocks in understanding. The gene for Huntington's disease has been known for a decade. The location of the protein is known, and a specific alteration is sufficient to cause the disease. However, we do not yet know why this causes the pathology of the disease^{25,26}. One of the chromosomal regions associated with inflammatory bowel disease is 5q31 (ref. 27). Intense genetic analysis using both linkage and population association studies has narrowed the critical interval to 250 kilobases (kb). Known sequence variants (many of which are tightly associated with the disease) have been examined in detail as possible risk factors²⁸, and yet we need more information before we can establish the mechanism of inflammatory bowel disease from this genetic association. In diabetes, dominant-negative mutations in the gene encoding the nuclear receptor peroxisome-proliferator-activated receptor- γ (PPAR- γ) are associated with severe insulin resistance, and the antidiabetic thiazolidinedione drugs have been shown to bind and activate the receptor²⁹. However, despite compelling genetic evidence of the involvement of PPAR- γ in glucose homeostasis, the mechanism by which the gene contributes to insulin sensitivity and glucose homeostasis is not understood. We cannot use the information we have on PPAR- γ to develop a better understanding of the role of adipocyte differentiation in diabetes or to find new targets for intervention. How can we gain new information systematically, improve our knowledge of the genome, and remove roadblocks like this?

Discovering all the functional information in the genome

The finite number of protein-coding genes hides a much greater diversity and extent of functional information in the human genome, most of which is still to be discovered. For example, alternative splicing allows multiple functions encoded by the same gene to be selected in a cell-specific manner (see ref. 30, for example). Multiple promoters can confer a diversity of inducible responses and substrate specificities on the same gene (see ref. 31, for example). Annotation of this degree of coding diversity in human genes is at a very early stage. It should also be noted that, as yet, only two-thirds of the genes have a canonical structure with an open reading frame. Developing the annotation at these levels requires manual curation and more experimental data, particularly to find many of the 5' ends and translation initiation codons of the genes. We are becoming aware of the existence of new classes of RNA genes. In addition to the well-known ribosomal, transfer and small nuclear RNAs, evidence is accumulating for the existence of other RNAs, such as antisense RNAs and microRNAs^{32,33}, both of which might be involved in gene regulation. MicroRNAs are 19–25-nucleotide products formed by the cleavage of precursor hairpins and might influence the translational activity or stability of mRNAs^{34,35}. Antisense RNAs act mainly by disrupting translation after hybridization to the sense mRNA³⁶. These discoveries reveal a new complexity of expressed information that is encoded in the genome, the extent of which is unknown.

The use of genomic information by each cell is governed by the interaction of multiple proteins with regulatory sequences that act as signal processors. As a result, a response is initiated that takes into account all the information received from either inside or outside the

cell^{37,38}. When analysing genome sequence, it is much harder to recognize regulatory sequences than protein-coding sequences, because the rules are more complex and less obvious. Yet, like the protein-coding regions, many regulatory sequences have been conserved during evolution, allowing us to use information from other organisms to try to find these functionally important elements of the human genome. Gene regulation is also governed by modifications to the DNA sequence or epigenetic changes (see review in this issue by Egger *et al.*, page 457), and an important adjunct to the genome sequence is to study patterns of differential methylation of genomic DNA and how they affect gene expression³⁹.

Not everything is to do with genes; much of the information in the genome sequence must be considered in the larger context of the chromosomes. Human chromosomes undergo replication and segregation by an intricate and controlled process. In yeast and other organisms, replication origins are sequence specific. It is possible that human replication origins also contain functionally important DNA sequence motifs, but we cannot recognize them yet. Important regions such as centromeres lie in highly repetitive heterochromatin; thus, we have not yet sequenced through a human centromere or determined the exact sequences needed for proper chromosome segregation. It is not known whether sequence motifs are involved in chromosome pairing or chiasma formation and crossing over during meiosis. Finally, DNA sequences might be involved in the positioning of chromosomes in the three-dimensional space of the nucleus; sequences on different chromosomes might be brought near to each other, leading to translocations and disease⁴⁰. How will we discover all the functional information stored in the human genome sequence?

Comparing genome sequences

Comparing whole genome sequences between species will make an important contribution to high-quality annotation, although by definition it will only reveal features that have survived during evolution, as opposed to features that are species specific. The first comparison of two mammalian genomes (human and mouse) revealed that where the two sequences can be aligned there is more than 68% nucleotide identity⁶ — too much for a very precise definition of conserved motifs. However, by comparing the variable degree of conservation in the aligned sequences, it was possible to deduce that about 5% of the human genome seems to be under selection to conserve features in common with the mouse ('purifying' selection)⁶. As expected, these features include most of the protein-coding exons (1.5% of the genome) and additional untranslated regions of the genes (1%). It is a high priority to target the other 2.5% in the search for new functionally important sequences.

Adding genome sequences from more organisms will greatly improve the analysis. In a comparison of 1.8 million bases of finished sequence of human chromosome 7 and that of 11 other vertebrates (an analysis termed 'phylogenetic footprinting'), 228 'multi-species conserved sites' (MCSs) overlapped virtually all the protein-coding exons, whereas another 966 were non-exonic^{41,42}. Extrapolating this analysis to the whole genome would result in the detection of about 200,000 protein-coding exons (close to estimates based on other analyses) and 1.5 million vertebrate-specific MCSs to help in pinpointing possible regulatory elements that have been as well-conserved during evolution as the protein-coding sequences. To highlight conserved sequences that are specific to primates, a similar approach would examine the variation between multiple primate genome sequences as a function of the evolutionary distance between them ('phylogenetic shadowing'). A pilot study suggests that an analysis of about seven carefully selected primate genome sequences would be sufficient to detect conserved signatures for a majority of protein-coding exons and many putative regulatory elements, including examples that are exclusive to primates and the hominid lineage⁴³.

Not all functional sequences are likely to be conserved between genomes. Comparisons within a genome sequence offer a complementary approach to the problem. For example, transcription-factor-

binding motifs (typically 6–10 base pairs in length; <http://www.gene-regulation.com/pub/databases.html#transfac>) occur in abundance throughout the genome. Searching for single motifs alone is not very informative. However, some well-studied promoters and other regulatory elements contain multiple motifs in short windows of genome sequence (typically less than 400 bases), and these can be correlated with experimental evidence^{38,44} and added to the annotation.

Functional annotation

Experimental confirmation of protein-coding gene structures is relatively straightforward. Ascertaining the structure and function of each protein is more of a challenge. The three-dimensional structures of only 1,539 distinct human proteins have been determined experimentally so far (<http://www.rcsb.org/pdb>), and currently the function of only 6,000 human proteins is known with certainty. Although the gene sequence does not generally provide the answer, it serves as the platform for determining function. It is possible to start with a gene sequence, disrupt it and correlate it with the resulting phenotype; or to start with a disrupted phenotype and map the mutation back to the sequence ('forward' or 'reverse' genetics, respectively). Either approach relies on the nature of the disrupted phenotype to determine the underlying biochemical function, which is then assigned to the gene. We rely extensively on using experimentally tractable model organisms for functional annotation. Accurate alignment of their corresponding genome sequences is crucial for extrapolating the experimental findings from one organism to another.

Understanding biological function also requires characterization of the many interactions of proteins with each other and with other cellular components. Large-scale screens for protein–protein interactions use methods involving yeast two-hybrid systems or immunoprecipitation of protein complexes with antibodies against native or tagged proteins. These methods require considerable validation to detect false positive and negative results, but the available experience suggests that it is possible to detect interactions involved in key cellular functions⁴⁵.

Experimental approaches are also important in identifying and validating sequences involved in gene regulation or chromatin function. In particular, methods are becoming available to recognize regions of chromatin that are subject to histone modification or transcription factor binding, and to map them back to the genome sequence^{38,46}. As high-quality experimental data sets of this kind emerge, they will need to be integrated as part of the annotation of the human genome in linked databases (see below).

Human sequence variation

Sequencing individual human genomes provides the richest source of genetic information. Projects underway have begun by selecting genes, exons or promoters for targeted sequencing in depth. On average there are fewer variants in protein-coding than in other sequences, and they tend to have lower minor allele frequencies. This presumably reflects selection against some variants that are non-synonymous (that is, those that cause amino-acid changes). If an excess in the rate of non-synonymous over synonymous substitution is observed in a particular coding region, it can be taken as an indicator of positive (diversifying) selection. As we find more variation and improve the sequence annotation, a fuller picture will emerge of the variants that alter genome function, and hence those that contribute to health and disease.

To realize the benefits of genomic medicine fully, we should start resequencing individual genomes in their entirety. With the benefit of emerging new technologies (reviewed in refs 47, 48), it is feasible to consider generating gigabases of data as short sequence reads (for example, 25–50 bases each) in a single experiment, and assembling the data accurately using the existing finished sequence as a template. The first 200 haploid human genome sequences will give us an in-depth view of human sequence variation (with a 99% chance of

detecting variants at an allele frequency of at least 1%; see table 2 in ref. 15). Comparison of these data sets will give us a full profile of common germline variation along each chromosome — something that is impossible to see with any of the current resequencing programmes. Detecting every variant (subject to the sensitivity and accuracy of the method) would permit the precise estimation of recombination rates and correlation with variability along chromosomes, and would give an indication of the forces of natural selection acting

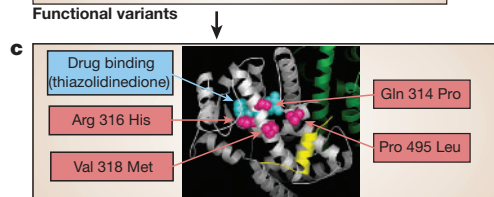
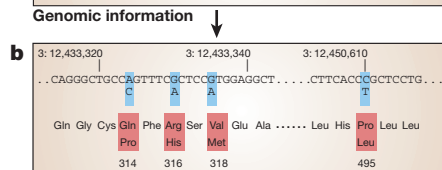
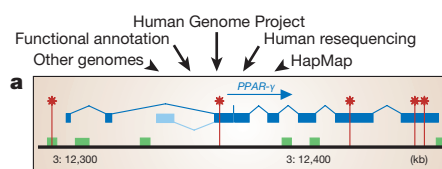
on the genome. This approach could be applied to give important baseline information in healthy tissue to compare with cancer genome sequences, and hence the ability to monitor DNA changes on a genome-wide basis during tumour progression. The same complete data sets could be generated for genomes that are associated with any other disease, with the prospect of removing the bias in ascertainment that is currently a limitation of gene-specific and regional investigations.

Box 1

Translation of genomic information to future clinical practice

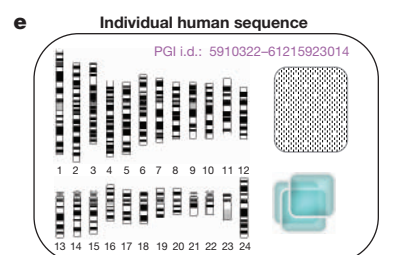
As the annotation of the human genome becomes stable, a user-friendly, distilled view can be developed, as in the figure. The diagram (a) of a chromosome 3 region (12,300–12,450 kb, numbering as in build 34, <http://www.ensembl.org>) contains the *PPAR-γ* gene structure (dark blue) with an alternative promoter (light blue), hypothetical noncoding functional regions (green shaded boxes), and functional variants (red). Note that introns in the gene structure are scaled down relative to the exons. Zooming in on two sequence segments (b) shows the translated sequence with functional variants highlighted in blue (nucleotide changes) and pink (amino-acid changes). Amino-acid numbering includes the propeptide sequence. The variants (c, pink) can be viewed in the monomer protein structure (grey) in a linked database. Also shown is the binding position of an antidiabetic thiazolidinedione drug (blue), part of the other monomeric unit (green) of the dimeric receptor, and the ligand (yellow). Using linked information from a range of sources, a summary of the known, modelled or predicted biological consequences (such as biochemical, structural, medical or pharmacological) could be curated (and updated regularly) for each functional variant in tabular form (d). A small subset of this information would define the disease or drug outcome or side effect associated with each variant, would constitute specific risk information of value in clinical assessment, and would be exported (red outlined boxes). For maximum usefulness, therefore, the exported information would be subject to stringent filters and would include only data for which the medical relevance was well established for each particular disease discipline. For example, variants of uncertain significance would be excluded from the filtered risk information, although all data would be available in the public domain. All the information in a–d would be curated in the public domain.

The use of personal genetic information in a clinical setting would be initiated or consented to by an individual. The individual sequence acquired could be as little as one or more individual genotypes, or as much as a complete genome sequence. The information would be private and owned by the individual, and

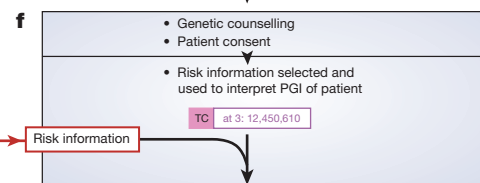


Genome base	3: 12,450,610
Genotype	t/c
Individual information	Pro/Leu
Biochemical consequence	
Structural consequence	Modelled: slight altered conformation of peptide backbone; increased local hydrophobicity.
Medical consequence	Known: associated with severe insulin resistance, diabetes mellitus and hypertension.
Pharmacological consequence	Known: resistant to thiazolidinediones.

Biological consequences

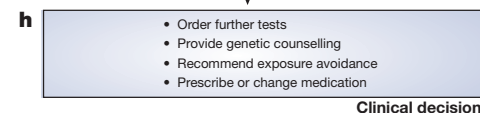


Personal genetic information (owned by individual)



PGI i.d.: 5910322-61215923014				
Disease: Diabetes (type2)				
Nucleotide position	Risk genotype	Individual risk	Medical risk	Other risks/effects
3: 12,450,610	TC		INS-resistance	Thiazolidinedione resistance
6: 149,031,974	GG	...		

Personal genetic assessment



Clinical decision

might be stored electronically, protected by a high-security code requiring unique personal identifiers (such as multiple fingerprint identification) for access only with consent of the individual (e). The information might be taken either before consultation (as illustrated here) or afterwards, and in either case would be subject to counselling by the practitioner and consent by the individual.

A specific investigation would be initiated by a consultation (f). The personal genetic information would then be supplied by the individual, for interpretation with respect to an agreed set of variants and/or a specific phenotype. The practitioner would use the available risk information concerning each variant to provide a genetic assessment for the individual (g). The top line refers to the variant featured in d and f; the second line is a hypothetical entry for a variant on another chromosome and does not represent a known variant. In the case illustrated, the individual has the heterozygous genotype TC at position 3: 12,450,610. This corresponds to having both Pro 495 and Ala 495 forms of the protein PPAR-γ. This genotype confers an increased risk of insulin-resistant diabetes on the individual, and also resistance to the thiazolidinedione class of antidiabetic drugs²⁹. Combining this with risk information for other genotypes would help to inform subsequent clinical decisions (h).

From genotype to phenotype

Observation of medical phenotypes is at the heart of accurate diagnosis. In addition to following rigorously defined clinical criteria, measurement of biochemical phenotype data helps considerably. For example, in a prospective cohort, 1% of children tested positive for antibodies against transglutaminase, a sensitive indicator for coeliac disease, although nearly all were asymptomatic with respect to gluten sensitivity. The disease incidence in adults is also 1% (ref. 49), indicating that we should revise our view of the onset of the condition, adopt the new phenotype data for genotype–phenotype association studies, and try to develop early preventive measures for those at risk. Accurate measurement of molecular variables can also be used to reclassify the disease ('molecular taxonomy') with greater phenotypic accuracy, as exemplified in the use of expression microarrays to subdivide some cancer types in retrospectively collected cases⁵⁰. These principles should be applied rigorously in characterizing population sample collections to underpin genotype–phenotype studies.

To find the major genetic factors in common disease, it is best to base the study on retrospective sample collections. This is the only way of obtaining enough cases of a well-defined phenotype to provide sufficient statistical power for the study. By contrast, prospective collection within the general population is appropriate for an unbiased sampling and measurement of environmental exposures before the onset of a specific condition. The challenge is to collect sufficient phenotype and exposure data on enough samples for statistically significant correlations. Studying environmental factors will enable us to explain more of the variance of common disease. Some genetic effects will be detected only in conjunction with environmental exposures. We can use genetic analysis to identify a susceptible or resistant subpopulation so as to strengthen the power to detect association with the environmental factors. We can learn how to provide individualized prevention advice based on genotypes. Correlations can be made on the basis of systematically collected, accurate data on factors such as nutrition, environmental toxins, exercise and lifestyle, with detailed information on individual genetic make-up. This would provide information for identifying avoidable risk, protective effects (either individually or on a population basis) for improved awareness, and recommendations for the implementation of health standards, food production, waste exposure and so on. Examples of regional and national programmes include the Avon Longitudinal Study of Parents and Children (<http://www.alspac.bris.ac.uk>), the Framingham study (<http://www.nhlbi.nih.gov/about/framingham>), the Estonia Genome Project (<http://www.geenivaramu.ee>), and the deCODE programme in Iceland (<http://www.decode.com>), an isolated population that arose from relatively few founders that exhibits elevated levels of association of certain rare phenotypes. Studies of large outbred populations, such as those in the United Kingdom and the United States, are also under development (<http://www.ukbiobank.ac.uk>; see commentary in this issue by Collins, page 475). There is a tremendous opportunity for these projects to stimulate public engagement at a new level. This is essential if these projects are to survive in the long term and if the general public are to embrace the new era of genetics.

Seeing is believing

The need to view, analyse and download genomic and functional data presents a formidable challenge. We need to have access to an enormous, complex and continually evolving body of information. At the same time, the displays must be flexible and user-friendly, allowing appropriate subsets of the data to be viewed clearly from any perspective (Box 1). For some model organisms, such as worm, fly and yeast, there is a single comprehensive database, but this seems to be impracticable for humans, given the scale and complexity of the data sets. Instead, there are genome 'browsers' at several sites, which gather data from multiple locations and layer it on the human genome sequence (see <http://genome.ucsc.edu>; <http://www.ensembl.org>;

<http://ncbi.nlm.nih.gov>). Personal data sets can be added to the display as separate 'tracks' and viewed either publicly or privately anywhere in the world by using DAS, the distributed annotation system⁵¹.

How will this evolve? At present, views of genome annotation include the supporting raw data—for example, individual expressed sequence tags, complementary DNA sequences, and protein homology matches can all be viewed underneath a gene structure that is computed automatically. This allows the individual user to judge the strength of the automatic prediction before using the information in further research. However, the automatic annotation inevitably includes unresolved conflicts, incomplete gene structures, missing data and errors. Two things are needed to make the genome more accessible: a stable, reliable core annotation, and a simple, distilled view of the genome for easy use in medical (and other) situations. Curated gene structures are already part of the annotation of the published human chromosomes (see refs 17, 52–59 and articles in this issue of *Nature* by Deloukas *et al.* (page 375) and Humphray *et al.* (page 369)) and are being displayed at a single dedicated site (<http://vega.sanger.ac.uk>). This should replace the automated reprocessing system, which would otherwise overwrite the results of the manual curation. Stable gene structures and similar information should become core annotation that is adopted as a single 'gold standard'. The core annotation should expand to include noncoding RNA genes, alternative transcripts, promoters and regulatory elements. This will be complemented by links to the specialist data sets (such as protein structures), which are best maintained by expert groups at local sites.

Medical sequencing

Medicine already benefits from the human genome sequence. Using the genome in basic research to help understand the cause of disease and variable response to toxins will be its most important application to medicine. The emphasis is shifting from positional cloning to screening candidate genes for disease-related variants. If necessary, it will soon be possible to scale up the process to cover the whole genome (or all genes), thus obviating the need for a prior hypothesis to select candidate genes or regions. New genes offer new avenues for diagnosis and intervention, and possibilities for translating the output of research into tangible improvements in health care. The other articles in this issue explore some of the current directions of genomic medicine.

The most striking future clinical application arising from the reference human genome sequence is the possibility of unlimited medical sequencing. For some years we have used targeted sequencing in diagnostic applications. For example, testing for mutations in specific genes or exons is used to assist genetic counselling in diseases such as cystic fibrosis or breast cancer. With easy access to a well-annotated human genome and cheap, accurate technology for whole-genome sequencing, an individual could acquire either a specific or a complete personal genetic health profile, including risk and resistance factors. This information could be used to improve and guide important medical decisions, to assess the risk of possible future exposures, and to select preventive treatments for improved health (Box 1). It would also provide important baseline information for retesting later in life. Consider the possibility of obtaining a full sequence on each case in the growing number of asbestos-induced mesotheliomas⁶⁰, finding somatic mutations in a key gene, and using the gene product as a target for the development of new anticancer drugs. Similarly, sequencing pathogen-resistant human genomes would provide new clues for countering our susceptibility to infectious disease, our biggest health burden worldwide. The use of new individual sequence information anonymously for research purposes (as in the last two examples above) should be relatively free of ethical constraints. However, it must be acknowledged that the increasing use of personal genetic information for individual health applications raises substantial ethical and moral issues. Personal information must remain the property of the individual, shared in

confidence only if proper consent has been given, and then with qualified practitioners. New levels of genetic counselling must accompany our practical ability to use the data so that, for example, risk information is used appropriately and is limited to guiding decisions for which there is a potentially beneficial treatment or action available.

The human genome sequence was made freely available for everyone to use. However, putting the genome sequence in the public domain also provides a golden opportunity for all scientists to contribute their expertise and vision to the genome. Only by continuing to make all the new data available will we develop a full understanding of the genome and provide a fertile environment for future innovation. □

doi:10.1038/nature02622

1. Bentley, D. R. Genomic sequence information should be released immediately and freely in the public domain. *Science* **274**, 533–534 (1996).
2. Guyer, M. Statement on the rapid release of genomic DNA sequence. *Genome Res.* **8**, 413 (1998).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
5. Rogers, J. The finished sequence of *Homo sapiens*. *Cold Spring Harb. Symp. Quant. Biol.* **68** (in the press).
6. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. Rat Genome Project Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
8. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
9. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
10. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
11. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474 (1997).
12. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
13. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
14. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
15. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
16. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
17. Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428**, 522–528 (2004).
18. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
19. Druker, B. J. Imatinib alone and in combination for chronic myeloid leukemia. *Semin. Hematol.* **40**, 50–58 (2003).
20. Aiuti, A., Ficara, F., Cattaneo, F., Bordignon, C. & Roncarolo, M. G. Gene therapy for adenosine deaminase deficiency. *Curr. Opin. Allergy Clin. Immunol.* **3**, 461–466 (2003).
21. Hacein-Bey-Abina, S. *et al.* A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **348**, 255–256 (2003).
22. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
23. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CCR5* structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273**, 1856–1862 (1996).
24. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
25. Van Dellen, A. & Hannan, A. J. Genetic and environmental factors in the pathogenesis of Huntington's disease. *Neurogenetics* **5**, 9–17 (2004).
26. Georgiou-Karistianis, N. *et al.* Future directions in research with presymptomatic individuals carrying the gene for Huntington's disease. *Brain Res. Bull.* **59**, 331–338 (2003).
27. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
28. Peltekova, V. D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nature Genet.* **36**, 471–475 (2004).
29. Barroso, I. *et al.* Dominant negative mutations in human PPAR γ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402**, 880–883 (1999).
30. Orr-Urtreger, A. *et al.* Developmental localization of the splicing alternatives of fibroblast growth factor receptor-2 (FGFR2). *Dev. Biol.* **158**, 475–486 (1993).
31. Gong, Q. H. *et al.* Thirteen UDP glucuronosyltransferase genes are encoded at the human *UGT1* gene complex locus. *Pharmacogenetics* **11**, 357–368 (2001).
32. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
33. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).
34. Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA* **101**, 2999–3004 (2004).
35. Ambros, V. *et al.* A uniform system for microRNA annotation. *RNA* **9**, 277–279 (2003).
36. Lehner, B., Williams, G., Campbell, R. D. & Sanderson, C. M. Antisense transcripts in the human genome. *Trends Genet.* **18**, 63–65 (2002).
37. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
38. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
39. Novik, K. L. *et al.* Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* **4**, 111–128 (2002).
40. Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A. & Misteli, T. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nature Genet.* **34**, 287–291 (2003).
41. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
42. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
43. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
44. Pilpel, Y., Sudarsanam, P. & Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **29**, 153–159 (2001).
45. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
46. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nature Biotechnol.* **21**, 1337–1342 (2003).
47. Smith, T. Whole genome variation analysis using single molecule sequencing. *Targets* (in the press).
48. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.* **5**, 335–344 (2004).
49. Bingley, P. J. *et al.* Undiagnosed coeliac disease at age seven: population based prospective birth cohort study. *Br. Med. J.* **328**, 322–323 (2004).
50. Golub, T. R. Genomic approaches to the pathogenesis of hematologic malignancy. *Curr. Opin. Hematol.* **8**, 252–261 (2001).
51. Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. The distributed annotation system. *BMC Bioinformatics* **2**, 7 (2001).
52. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
53. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
54. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
55. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
56. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
57. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
58. Skaltsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
59. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).
60. Treasure, T., Waller, D., Swift, S. & Peto, J. Radical surgery for mesothelioma. *Br. Med. J.* **328**, 237–238 (2004).

Acknowledgements I thank I. Barroso, A. Coffey, T. Cox, S. Grant, T. Hubbard, S. Hunt, G. Leschziner, E. Margulies, K. Rice, J. Rogers, M. Ross, C. Shaw-Smith, R. Steward, M. Stratton, C. Tyler-Smith and others for assistance, discussion and critical reading of the manuscript. The author is supported financially by the Wellcome Trust.

Competing interests statement The author declares that he has no competing financial interests.

Mapping complex disease loci in whole-genome association studies

Christopher S. Carlson¹, Michael A. Eberle³, Leonid Kruglyak^{2,3} & Deborah A. Nickerson¹

¹Department of Genome Sciences, University of Washington, 1705 NE Pacific, Seattle, Washington 98195-7730, USA

(e-mail: csc47@u.washington.edu)

²Howard Hughes Medical Institute and ³Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA

Identification of the genetic polymorphisms that contribute to susceptibility for common diseases such as type 2 diabetes and schizophrenia will aid in the development of diagnostics and therapeutics. Previous studies have focused on the technique of genetic linkage, but new technologies and experimental resources make whole-genome association studies more feasible. Association studies of this type have good prospects for dissecting the genetics of common disease, but they currently face a number of challenges, including problems with multiple testing and study design, definition of intermediate phenotypes and interaction between polymorphisms.

After decades of study, much has been learned about the genetics of common diseases, but most of this knowledge relates to rare families segregating high-risk alleles¹. Such alleles are generally very rare in the population and therefore explain relatively little of overall disease prevalence. A simple way of describing the relative importance of a locus from the standpoint of public health is to use the population attributable fraction (PAF). This factor can be thought of as the fraction of the disease that would be eliminated if the risk factor were removed. In a public health context, high-risk alleles may have a large PAF (>50%) in rare, single-gene mendelian diseases such as cystic fibrosis^{2,3}, but do not appear to be important for common diseases. The PAF for rare alleles in common disease is generally less than 10%. For example, more than 150 rare high-risk alleles have been identified for Alzheimer's disease in three genes (*PS1*, *PS2* and *APP*; for more information, see <http://molgen-www.uia.ac.be/ADMutations>), but the combined PAF for all of these alleles is less than 5% of all Alzheimer's cases⁴.

Common modest-risk alleles may account for a greater PAF in common disease than do rare high-risk alleles; this is often referred to as the common disease/common variant (CDCV) hypothesis⁵. For example, the PAF for the Apoε4 allele in late-onset Alzheimer's disease has been estimated at 20% (ref. 6), which is greater than all of the known rare high-risk alleles combined.

Scientists are motivated by three factors in their search for common modest-risk variants associated with common disease: identified risk variants can illuminate new and important aspects of disease pathogenesis; common variants can be more important than rare ones from a public health perspective because they usually have a larger associated PAF; and common modest-risk variants are easier to identify than rare modest-risk variants⁷, as will be discussed later. Association studies compare the allele frequency of a polymorphic marker, or a set of markers, in unrelated patients (cases) and healthy controls to identify markers that differ significantly between the two groups. Many papers have been published detailing studies using just one or a small number of polymorphisms, but many, if not most, of the identified associations have been difficult to replicate⁸. Because it seems likely that common modest-

risk variants exist^{8,9}, developing tools such as high-density genetic maps for whole-genome association analyses is an important goal in the next phase of analysing the human genome. Comprehensive analysis of candidate regions, or even the whole genome, should produce more robust results. Over the past several years, the technical¹⁰, informatic¹¹ and statistical¹² foundations have been laid for whole-genome association analyses^{7,13–16}. It is important that researchers planning to use these resources understand that association analysis using dense maps, such as the HapMap¹⁷, is fundamentally different from the family linkage studies used to identify rare high-risk alleles.

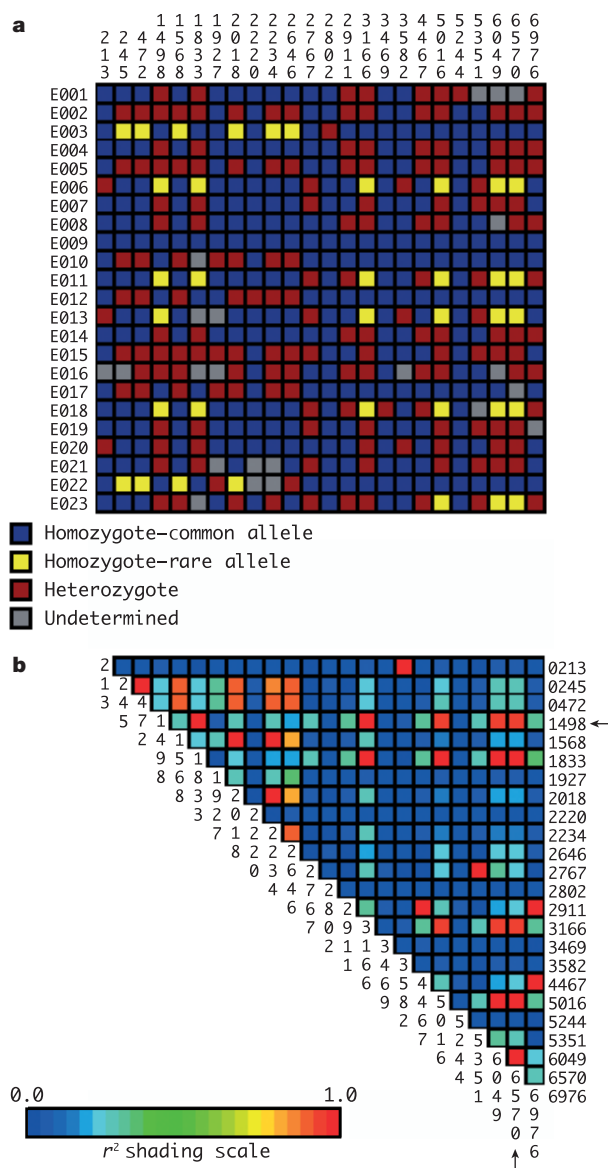
In this review, we will address the prospects for whole-genome association studies, starting with a summary of the theoretical basis of association analysis and currently available experimental resources. We will compare the strengths and weaknesses of direct, indirect and pooled study designs, and discuss the prospects for application of these designs at a whole-genome level. We close with a discussion of the challenges associated with multiple testing in whole-genome association analysis, both at the level of single polymorphisms and at the level of interaction between polymorphisms.

Linkage versus association

For the past two decades, the dominant study design for investigation of the genetic basis of inherited disease has been linkage analysis in families. Linkage analyses search for regions of the genome with a higher-than-expected number of shared alleles among affected individuals within a family. This indicates that somewhere within this linked region there is a disease-predisposing allele. Closely related individuals tend to share large regions of the genome inherited from the same recent ancestor, and therefore genotyping fewer than 500 polymorphic markers across the genome is generally adequate to detect linked regions. In linkage analysis, any polymorphism between a pair of linked markers will be associated with both markers. As a result, linkage maps have a logical hierarchical structure wherein an initial genome scan can be performed at low density (fewer than 500 markers), with additional markers used to fine-map the boundaries of linked regions. Because of the small number of recombination events within most families, it is frequently difficult to narrow the region of

Box 1

The non-hierarchical structure of linkage disequilibrium



In an indirect association analysis, a dense set of markers across a region is genotyped and tested for association. The assumption is that the genotype at polymorphisms in the region that are not genotyped directly will be correlated with one or more of the markers. The strength of the correlation between markers and an untyped polymorphism (frequently referred to as linkage disequilibrium, LD) can be described using several statistics. From the standpoint of association analysis, the most appropriate LD statistic is r^2 , the Pearson correlation coefficient between alleles at each locus²⁵. r^2 is directly related to statistical power to detect an untyped SNP. The ability to detect a risk SNP indirectly in n samples is roughly equivalent to the ability to detect it directly in nr^2 samples⁶¹. Thus, even in a large number of samples, low r^2 between markers and adjacent unassayed polymorphisms will result in low power to detect disease associations at the unassayed polymorphisms.

As an example, a visual genotype for the common SNPs in the *IL10* gene in a European population is shown in **a**, with r^2 between SNPs shown in **b** (image generated with the VG2 program; <http://pga.gs.washington.edu/VG2.html>). The important point is that even in a 7-kilobase, non-recombinant region (sometimes referred to as a 'haplotype block'), there are clearly several common patterns of variation, and levels of r^2 between these patterns are low. If we genotyped SNPs 1498 and 6570 as markers (indicated by arrows), we would observe strong LD between them ($r^2 = 1$; shown in red). But there are clearly a number of SNPs between these two with distinct patterns of genotypes (for example, 2767, 2911, 3582 and 4467) that are weakly associated with either marker and therefore would not be detected. Careful scrutiny will reveal that each common pattern is seen at least twice, effectively ruling out recurrent mutation as a source of variation in the region. This graphically illustrates how LD between adjacent markers that are not associated with a phenotype cannot be used to hierarchically exclude the intervening region from containing a disease-associated SNP.

interest to below several megabases, but candidate-gene analyses within linked regions have been greatly accelerated by the availability of the complete sequence of the human genome.

Linkage analysis is more powerful than association analysis for identifying rare high-risk disease alleles, but association analysis is expected to be more powerful for the detection of common disease alleles that confer modest disease risks⁷. This reflects the fact that for modest-risk alleles the patterns of allele sharing among affected individuals within pedigrees are less striking than patterns of allele sharing between unrelated affected individuals. Another advantage of association analysis is that it is easier to recruit large numbers of unrelated affected individuals than it is to collect large numbers of pedigrees, each containing multiple affected individuals, especially for diseases of old age. But the region around a marker that is shared identically by descent in unrelated, affected individuals will be much smaller than the shared region for related individuals because of the much higher number of generations from the most recent common ancestor. Therefore, association analysis requires markedly higher marker densities than linkage analysis, of the order of hundreds of thousands of markers^{13,17}.

Single-nucleotide polymorphisms (SNPs) are variations in DNA sequence where one of the four nucleotides is substituted for another (for example, C for A). SNPs are the most frequent type of polymorphism in the genome, and therefore will make up the vast majority of markers in a whole-genome association map. We will refer to SNPs genotyped in an association analysis as tagSNPs; a wide variety of techniques for selecting tagSNPs exist^{10,18–23}, with many more under development.

Linkage disequilibrium (LD) describes the non-random correlation between alleles at a pair of SNPs. A major practical difference between linkage maps and association maps is that LD does not have a hierarchical structure akin to linkage¹². In association analysis, a SNP between a pair of strongly associated tagSNPs may not be strongly associated with either tagSNP (Box 1) and therefore can easily go undetected. Association analysis can identify disease-risk variants only when such variants are strongly associated with a tagSNP.

The first generation of whole-genome association maps will be available shortly, but as denser maps become available it will be necessary to genotype additional tagSNPs across the entire genome until

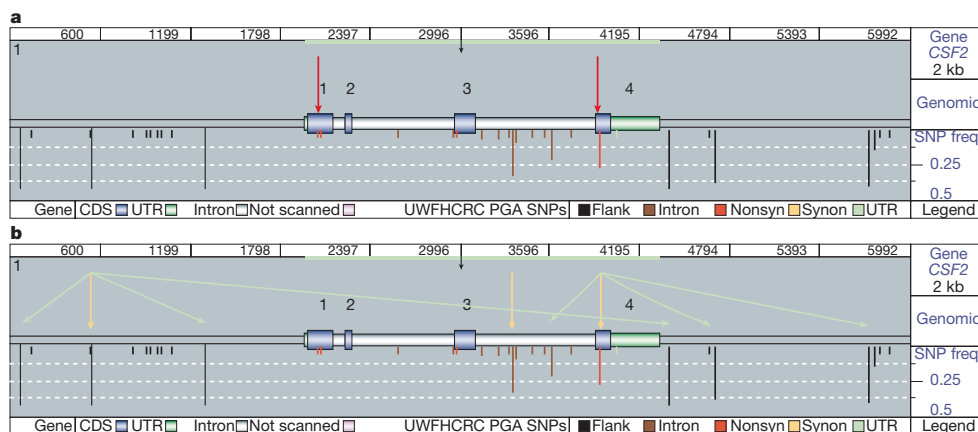


Figure 1 Direct versus indirect association analysis. **a**, In direct association analysis, all functional variants (red arrows) are catalogued and tested for association with disease. A GeneSNPs image of the *CSF2* gene is shown. Genomic features are shown as boxes along the horizontal axis (for example, blue boxes indicate exons). Polymorphisms are shown as vertical bars below the axis, with the length of the line indicating allele frequency and colour indicating context (for example, red indicates

coding SNPs that change amino acids). **b**, For indirect association analysis, all common SNPs are tested for function by assaying a subset of tagSNPs in each gene (yellow arrows), such that all unassayed SNPs (green arrows) are correlated with one or more tagSNPs. Effects at unassayed SNPs (green arrows) would be detected through linkage disequilibrium with tagSNPs. Images adapted from GeneSNPs (<http://www.genome.utah.edu/genesnps>).

it becomes clear that most of the newly typed tagSNPs are strongly associated with tagSNPs that have previously been genotyped. Once patterns of LD between common polymorphisms have been described, it will be possible to select optimized subsets of tagSNPs such that all common variants (or a sufficiently large fraction thereof) are either directly assayed or are strongly correlated with either an allele of a single tagSNP or a combination of alleles at several tagSNPs (a tagSNP haplotype). The optimal subset will not necessarily be the same for all studies, but will reflect a number of practical considerations specific to the study, including the expected effect size and desired statistical power, the genetic diversity of the study population and the available budget. As genotyping and ultimately whole-genome sequencing becomes cheaper²⁴, map optimization may not be necessary at all.

Current applications of SNP maps

Linkage analysis generally identifies broad intervals of several megabases that correlate with disease status within pedigrees, and these intervals can encompass dozens to hundreds of candidate genes. If the disease-predisposing allele is relatively common, it should be possible to fine-map the disease allele within linkage peaks using LD²⁵. Where a significant linkage hit has been identified and there is evidence of a single major-risk allele (as is found in haemochromatosis and cystic fibrosis), this is a perfectly reasonable strategy. However, a large number of 'suggestive' linkages exist for which the underlying genetic defect (if it exists) has not been identified. Suggestive linkage describes genomic regions with an observed trend toward excess sharing in affected individuals that is not significant after correcting for multiple tests across the genome^{26,27}. Suggestive-linkage regions can contain common disease alleles with modest relative risk, so there is some enthusiasm for association analysis of candidate regions of the genome identified as suggestive linkages. These linkages can also be attributed to simple false-positive results, with one suggestive-linkage hit expected per linkage scan under the null hypothesis that no risk allele exists. Thus, fine-mapping of suggestive-linkage peaks by association analysis makes sense only when simulations show clear evidence for an excess of suggestive linkage across the genome²⁸, or when the cost of doing so is acceptable without compelling evidence for linkage.

Candidate SNP analysis of coding SNPs (cSNPs) that change amino acids has intriguing potential, because the number of common cSNPs is several orders of magnitude smaller than the number of common SNPs overall. In resequencing 262 genes in the PDR90 panel (which consists of 90 human DNA samples with diverse ethnic origins²⁹), we have identified 147 cSNPs with an allele frequency above 5% (unpublished data). Extrapolating to 30,000 genes, a comprehensive analysis of common cSNPs could require ~20,000 common cSNPs to cover the genome, consistent with previous reports^{30–32}. This approach is currently impractical because there is no concerted public effort to comprehensively identify cSNPs, but we believe that such a resource would be very valuable and should be made a high priority of the genome project.

cSNPs constitute the majority of disease alleles in mendelian disorders, and a recent analysis³ suggests that common disease variants are likely to show a similar trend. Mendelian diseases are largely recessive, with mutations that abolish the normal function of a gene, as demonstrated by the relatively high frequency of nonsense mutations, so if common disease alleles reflect modest changes in gene function or activity, then the mendelian alleles are an inappropriate model. Furthermore, recent comparative genomic studies have demonstrated that the level of evolutionarily conserved non-coding sequence is comparable to the amount of evolutionarily conserved exonic sequence^{33,34}. It seems quite plausible that disease-associated variants with modest effect will be distributed proportionately between noncoding and coding sequences. However, our ability to identify functional variation in conserved NCSs is still in its infancy^{35,36}.

Candidate SNP analysis is a direct test of association between a putatively functional variant and disease risk. The alternative, which we will refer to as indirect association (Fig. 1), is to test a dense map of SNPs for disease association under the assumption that if a risk polymorphism exists it will either be genotyped directly or be in strong LD with one of the genotyped tagSNPs^{5,13,18}. The advantage of indirect association analysis is that it does not require prior determination of which SNP might be functionally important, but the disadvantage is that a much larger number of SNPs needs to be genotyped. Both direct and indirect association testing currently can be applied effectively to candidate genes that have been implicated in disease pathogenesis by other means, as long as common variants

have been comprehensively identified in the candidate gene, and the two approaches are not mutually exclusive.

Measured allele frequencies in pooled samples can be used in association analysis instead of genotyping individual samples^{37–39}. Using quantitative genotyping technologies to measure allelic frequency changes between pools of patient and control DNAs is considerably cheaper than genotyping individual samples, even though multiple replicates are required to increase the precision of allele frequency estimates in pools. Because pooling is cheaper, many more SNPs can be screened for association in a pooled analysis. Pooled studies will no doubt find some associations^{41–43}, but will probably miss others for one of two reasons.

The power of pooled analysis depends on the expected effect size: large relative risks can lead to large differences in allele frequency (Δp) between cases and controls, whereas small relative risks (for example, at modest-risk alleles) produce smaller differences. The ability to detect a real allele frequency difference is therefore a function of the true Δp between cases and controls and the error in estimating Δp . Error in Δp estimates can be divided into sampling error (random noise introduced by estimating Δp in a finite sample) and measurement error (noise introduced by the precision of the technological platform used to estimate Δp). Whereas sampling error decreases with increasing sample size, measurement error does not, and seemingly small measurement errors in Δp can dramatically reduce the power to detect modest-risk alleles. This is especially true in whole-genome studies, where the large number of tests virtually guarantees that true positives will be lost in a sea of false positives. For example, the expected Δp is just 4.3% for a risk allele with 10% allele frequency and 1.5-fold genotype relative risk (GRR; the relative risk associated with each risk allele carried by an individual), which could easily be missed in a whole-genome scan with measurement error of $\pm 2\%$, as observed in most pooling platforms⁴⁴.

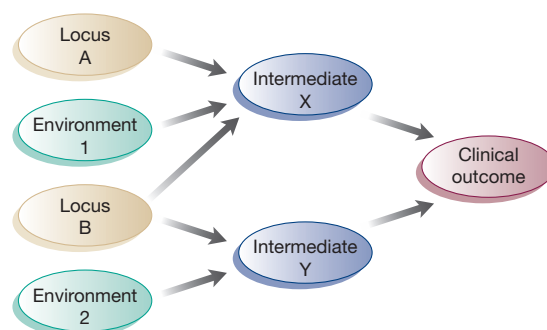
Technical considerations aside, pooled DNA analysis can be considerably less powerful than individual genotype analysis when known risk factors have been measured in each sample. This is true because common diseases generally have known environmental and genetic risk factors (for example, smoking status or sex). Analyses corrected for known risk factors should have greater power to detect novel factors when the risk associated with an unidentified genetic variant is smaller than that associated with a known factor, or if there is an interaction between the known risk factor and the unidentified genetic variant. This shortfall can be overcome by sorting cases and controls into subpools on the basis of known risk factors (such as smoking and sex, or genotype at a known risk locus) before the pools are genotyped. However, subpooling also increases the number of assays per SNP, so the economic advantage over individual genotyping is less marked. Even if subpools are assembled on the basis of known risk factors, the pooled frequency data cannot be used to explore novel gene–gene or gene–environment risk interactions.

The joy of intermediate phenotypes

Clinical outcome can be thought of as a synthesis (or grand total) of many risk factors, with intermediate phenotypes as subtotals. An example of this would be lipid profiles and risk of myocardial infarction, where the number of variables affecting low-density lipoprotein (LDL) levels is presumably smaller than the number of variables affecting the risk of myocardial infarction. Genetic factors contributing to intermediate phenotypes will generally be easier to identify because of the improved signal-to-noise ratio in the fraction of variance explained by any single factor (Box 2). Thus, it is crucial to appreciate how the definition of a phenotype can affect the prospects of an association analysis. Studies using a single clinical endpoint are akin to a shot at the moon with only one chance of success, whereas studies that collect multiple phenotypes are more likely to help us to understand the contribution of genetic factors to components of disease, regardless of whether a significant effect can be established on the clinical endpoint.

Box 2

Intermediate phenotype analysis



Intermediate clinical phenotypes have been identified for many diseases. Just as clinical outcome can be treated as a synthesis of the intermediate phenotypes, so intermediates can be treated as syntheses of subsets of proximal risk factors, both environmental (Environment 1 and 2) and genetic (Locus A and B), as shown in the figure. The advantage of intermediate phenotypes is that the number of genetic and environmental factors influencing each intermediate is presumably smaller than the number of factors affecting the clinical endpoint. Therefore, the proportion of variance in an intermediate phenotype explained by a given genetic locus will be greater than the proportion of variance explained in the clinical endpoint. A good example of this is the relationship between the gene *PON1* and risk of carotid artery disease (CAAD). The enzymatic activity of *PON1* is a known risk factor in CAAD⁶² and is a useful intermediate phenotype. Examination of genetic variants at the *PON1* locus has identified coding polymorphisms^{63,64} and regulatory SNPs that significantly alter *PON1* levels⁶⁵, but none of these SNPs showed significant associated risk for CAAD directly^{62,66}. In this case, the intermediate phenotype (enzyme activity) helped in the identification of a group of SNPs that probably do have a small effect on CAAD risk, but a risk increase that is so small that it could not be detected directly in the available sample.

When intermediates are relatively cheap (for example, standard clinical chemistry) they should be broadly applied, although multiple testing can become an issue because each additional phenotype increases the number of statistical tests. More expensive intermediate phenotypes, such as RNA expression and protein levels, or measures of drug response and metabolism, also deserve thoughtful consideration, and new high-throughput proteomic approaches may dramatically expand the feasibility of collecting intermediate phenotype data⁴⁵. Intermediate phenotypes may currently be a better investment than genotypes, at least until dense SNP-mapping technologies mature, especially considering that early studies will not necessarily help us to focus later efforts, because dense SNP maps do not have a hierarchical structure.

Although genotypes are reasonably precise, phenotypic and environmental risk factors are generally imprecisely measured, and the precision of these covariates will be very important in evaluating gene–environment interactions. For example, body mass index (BMI) is a more precise (and useful) measure than is a dichotomous obese/non-obese variable. For that matter, circulating levels of the nicotine metabolite cotinine are certainly a better measure of smoking status than self report⁴⁶. Just as intermediate phenotypes increase the probability of learning something useful from a study, a modest number of carefully phenotyped samples can be more valuable than a large number of poorly characterized samples. Conversely, studies that make a reasonable effort to regulate environmental exposures

should be better able to identify modest genetic effects because of reduced environmental noise.

Can the multiple testing problem be solved?

Before dense SNP maps are applied to search for genetic components of a complex disease, we need to understand how many markers will be required for a complete, high-power genome scan. More than seven million SNPs have been described in the human genome. The International HapMap Project¹⁷ aims to genotype at least 600,000 of these to define haplotype patterns across the genome, with the goal of developing a map of non-redundant tagSNPs, and is committed to denser spacing where necessary. Ultimately, such a map is likely to require 200,000 to one million markers to achieve a reasonable likelihood that any common SNP in the genome is usefully associated with at least one tagSNP¹⁷, which is consistent with previous theoretical estimates¹³. This generates an enormous multiple testing problem, and it will be difficult to achieve 'statistical significance' in a single genome-wide association study for all but the strongest effects. However, there are practical methods that may help to identify important genetic factors more efficiently, such as ranking markers by proximity to candidate genes or by expected functional consequence (for example, cSNPs).

Previous analyses have estimated the ability to detect a given GRR in the context of a case-control study. In one model of a genome-wide association study, a GRR of >1.5 associated with a common SNP (minor allele frequency of 10%) could be detected in a sample of ~2,000 matched case-control pairs⁷, conservatively corrected for one million independent tests. More recent analysis suggests that detecting a similar GRR would require many more samples, but for this the investigators considered a map of only 300,000 tagSNPs⁴⁷. These studies raise an important issue: even if it is difficult to establish significance within a single study, initial scans will at least help to focus subsequent efforts on interesting regions of the genome. Furthermore, these analyses conservatively assumed independent markers, but many markers in SNP maps will be significantly associated with one another. Thus, it is difficult to theoretically establish a threshold for significance in whole-genome association analysis, and significance will be easier to address empirically by permutation analysis of the observed data. False-discovery-rate analysis may also prove to be a useful tool in this area⁴⁸.

Higher-order risk interactions

One of the possible reasons that linkage analysis has identified few high-frequency, modest-risk alleles is statistical interaction (or epistasis) between multiple loci. If the risk associated with a given locus is influenced by the genotype at another locus, single-locus analyses might miss the association. Given that a large proportion of common disease risk is heritable and that the genetic risk factors identified by linkage explain relatively little of the PAF for common disease, epistatic interactions could well be important in predicting disease risk. If multiple testing is a major challenge at the level of single SNPs, the problem rapidly becomes intractable when we allow for gene-gene or gene-environment interactions. That is, for 600,000 tagSNPs and ten environmental variables, there are more than 10^{11} possible pairwise gene-gene interactions, and six million possible gene-environment interactions. Can we logically limit the search space for gene-gene and gene-environment interactions?

Overall disease risk can be modelled as the product of risks at many independent risk loci. With such a model, high-risk combinations of genotype will exist, but the ability to detect any single locus is a function of the relative risk of that locus alone. So what exactly is an interaction? Statistical interaction between loci requires a dependent effect, wherein the risk associated with a genotype at a locus is dependent on a genotype at another locus. There are several possible epistatic models. Assuming two risk loci (A and B), a synergistic epistatic effect occurs when the combined risk to risk-allele carriers at both loci is greater than would be

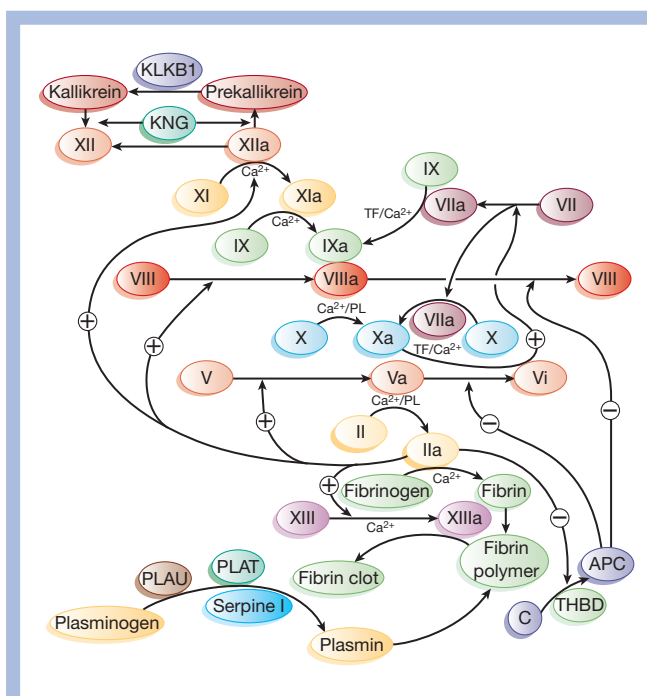


Figure 2 Pathway of physical interactions. If the risk associated with a given locus is influenced by another locus, single-locus effects may not be detected at either locus. Identifying interacting loci has improved with the introduction of expertly curated pathways like the one shown here for genes involved in blood clotting. Each gene is represented as an oval, with plain arrows representing conversion of proteins from one form to another (for example, VII to VIIa). Arrows with + or - symbols indicate enzymes that facilitate or inhibit conversions (for example, Xa converts VII to VIIa). Each type of epistatic risk can conceivably take place in such a pathway. For example, balanced risk interactions are plausible for physically interacting proteins (for example, factor IIa/fibrinogen), whereas permissive interactions are plausible for genes that are involved in the same pathway stream (for example, factors VIIa/II).

expected if they were independent, and an antagonistic epistatic effect is when the combined risk is lower than expected. A balanced epistatic effect is a special case of antagonistic effects where the risk for risk-allele carriers at both loci is actually lower than for risk-allele carriers at only one locus. Finally, a permissive epistatic interaction is when only risk-allele carriers at both loci show increased risk of disease.

Although relatively few epistatic interactions have been described to date⁴⁹⁻⁵², there is little doubt that many such interactions exist. But the sheer number of potential interactions practically guarantees that a comprehensive search has no power to detect them. In practice, there are several reasonable approaches to reduce the number of interactions analysed, such as limiting analysis to biologically plausible interactions between genes in related pathways, or limiting analysis to markers with appreciable single-locus effects. In general, risk interactions are more plausible between genes involved in a physical interaction, found in the same pathway or involved in the same regulatory network. Otherwise it is difficult to construct a model in which risk at one locus can be dependent on the genotype at the second locus. It would be reasonable to test for epistatic interactions between all pairs of non-synonymous cSNPs in physically interacting genes or pathways. However, identifying interacting genes requires improved bioinformatics tools to expertly annotate known pathways of physical interaction^{53,54} (Fig. 2), mine gene-ontology resources^{55,56} and annotate expression analyses for evidence of co-regulation⁵⁷. Fortunately, all of these are likely to become a reality in the near future.

Another approach to limiting the search space is to consider only interactions between SNPs that exhibit some evidence of a main or

single-locus effect. Epistatic risks with a large PAF are likely to exhibit large main effects, so it might be reasonable initially to restrict epistatic analysis to SNPs with large observed main effects. A test statistic conditioned on the observed marginal genotype frequencies at each SNP in cases could be used to identify epistatic interactions within cases, independently from the epistasis model or main effect at either locus. Similar tests for interaction have been described for gene–environment interactions, where the magnitude of an interactive effect can accurately be estimated from case data alone^{58–60}. Testing the 1,000 tagSNPs with the largest main effects for pairwise epistatic interaction would result in ~500,000 tests, comparable to the number of single-locus tests in a whole-genome scan. Conveniently, significant epistatic interactions among these tagSNPs can provide independent evidence that the main effects are real. As a final note on the exploration of epistatic interaction, the methodologies suggested above are solely an attempt to maximize power to identify statistically significant results within a single study. After the statistical power of a given study has been exhausted, it is still important to examine all possible interactions, although such a search is computationally daunting. Complete exploration of non-obvious interactions can help refine underlying analytical assumptions and study design, as well as generate testable hypotheses for subsequent efforts.

Future directions

The tools for the genetic dissection of complex traits are nearing maturity, and dense SNP maps will allow researchers to test the CDCV hypothesis and identify such variants where they exist. These discoveries will complement the knowledge already gleaned from rare mendelian disorders, providing information about which gene pathways are important in disease pathogenesis, which genes and alleles within these pathways are important risks in the general population, and how much of the heritable risk for common disease remains to be explained by rare modest-risk alleles and epistatic interactions. These tools may also facilitate the identification of important risk interactions, but considerable theoretical efforts to develop test statistics will be necessary before this can be achieved. Common variants identified will not wholly explain the genetic component of common disease risk, as there are certain to be rare modest-risk variants that will go undetected by these analyses, but the identification of important common variants will provide new targets for diagnostics, prognostics and therapeutics. □

doi:10.1038/nature02623

- Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (suppl.), 228–237 (2003).
- Grody, W. W. *et al.* PCR-based screening for cystic fibrosis carrier mutations in an ethnically diverse pregnant population. *Am. J. Hum. Genet.* **60**, 935–947 (1997).
- Kosorok, M. R., Wei, W. H. & Farrell, P. M. The incidence of cystic fibrosis. *Stat. Med.* **15**, 449–462 (1996).
- Rocchi, A., Pellegrini, S., Siciliano, G. & Murri, L. Causative and susceptibility genes for Alzheimer's disease: a review. *Brain Res. Bull.* **61**, 1–24 (2003).
- Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
- Slooter, A. J. *et al.* Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: the Rotterdam Study. *Arch. Neurol.* **55**, 964–968 (1998).
- Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
- Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
- Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
- Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5**, 89–100 (2004).
- Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
- Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* **71**, 1386–1394 (2002).

- Risch, N. & Teng, J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
- Risch, N. Evolving methods in genetic epidemiology. II. Genetic linkage from an epidemiologic perspective. *Epidemiol. Rev.* **19**, 24–32 (1997).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
- Ke, X. & Cardon, L. R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287–288 (2003).
- Stram, D. O. *et al.* Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36 (2003).
- Weale, M. E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73**, 551–565 (2003).
- Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
- Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
- Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
- Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
- Sawcer, S. *et al.* Empirical genomewide significance levels established by whole genome simulations. *Genet. Epidemiol.* **14**, 223–229 (1997).
- A meta-analysis of whole genome linkage screens in multiple sclerosis. *J. Neuroimmunol.* **143**, 39–46 (2003).
- Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
- Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
- Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
- Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
- Schwartz, S. *et al.* MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31**, 3518–3524 (2003).
- Mayor, C. *et al.* VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).
- Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
- Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
- Bansal, A. *et al.* Association testing by DNA pooling: an effective initial screen. *Proc. Natl Acad. Sci. USA* **99**, 16871–16874 (2002).
- Mohlke, K. L. *et al.* High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl Acad. Sci. USA* **99**, 16928–16933 (2002).
- Howell, W. M., Evans, P. R., Wilson, P. J., Cawley, M. I. & Smith, J. L. HLA class II DR, DQ, and DP restriction fragment length polymorphisms in rheumatoid arthritis. *Ann. Rheum. Dis.* **48**, 295–301 (1989).
- Yang, Y. *et al.* Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl Acad. Sci. USA* **100**, 7225–7230 (2003).
- Permutt, M. A. *et al.* Searching for type 2 diabetes genes on chromosome 20. *Diabetes* **51** (suppl. 3), S308–S315 (2002).
- Barcellos, L. F. & Thomson, G. Genetic analysis of multiple sclerosis in Europeans. *J. Neuroimmunol.* **143**, 1–6 (2003).
- Kammerer, S. *et al.* Amino acid variant in the kinase binding domain of dual-specific A kinase-anchoring protein 2: a disease susceptibility polymorphism. *Proc. Natl Acad. Sci. USA* **100**, 4066–4071 (2003).
- Sham, P., Bader, J. S., Craig, I., O'Donovan, M. & Owen, M. DNA pooling: A tool for large-scale association studies. *Nature Rev. Genet.* **3**, 862–871 (2002).
- Cristea, I. M., Gaskell, S. J. & Whetton, A. D. Proteomics techniques and their application to hematology. *Blood* **103**, 3624–3634 (2004).
- Haufron, V. & Lison, D. Urinary cotinine as a tobacco-smoke exposure index: a minireview. *Int. Arch. Occup. Environ. Health* **71**, 162–168 (1998).
- Shork, N. J. Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* **70**, 1480–1489 (2002).
- Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
- Bolk, S. *et al.* A human model for multigenic inheritance: phenotypic expression in Hirschsprung disease requires both the *RET* gene and a new 9q31 locus. *Proc. Natl Acad. Sci. USA* **97**, 268–273 (2000).
- Zetterberg, H., Zafiroopoulos, A., Spandios, D., A. Rymo, L. & Blennow, K. Gene–gene interaction between fetal MTHFR 677C>T and transcobalamin 776C>G polymorphisms in human spontaneous abortion. *Hum. Reprod.* **18**, 1948–1950 (2003).
- Butt, C. *et al.* Combined carrier status of prothrombin 20210A and factor XIII-A Leu34 alleles as a strong risk factor for myocardial infarction: evidence of a gene–gene interaction. *Blood* **101**, 3037–3041 (2003).
- Tiret, L. *et al.* Synergistic effects of angiotensin-converting enzyme and angiotensin-II type 1 receptor gene polymorphisms on risk of myocardial infarction. *Lancet* **344**, 910–913 (1994).
- Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.* **31**, 19–20 (2002).

54. Bonner, A. E., Lemon, W. J. & You, M. Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis. *J. Med. Genet.* **40**, 408–417 (2003).
55. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
56. Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
57. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
58. Khoury, M. J. & Flanders, W. D. Nontraditional epidemiologic approaches in the analysis of gene–environment interaction: case–control studies with no controls! *Am. J. Epidemiol.* **144**, 207–213 (1996).
59. Begg, C. B. & Zhang, Z. F. Statistical analysis of molecular epidemiology studies employing case–series. *Cancer Epidemiol. Biomarkers Prev.* **3**, 173–175 (1994).
60. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med.* **13**, 153–162 (1994).
61. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
62. Jarvik, G. P. *et al.* Paraoxonase (PON1) phenotype is a better predictor of vascular disease than is PON1(192) or PON1(55) genotype. *Arterioscler. Thromb. Vasc. Biol.* **20**, 2441–2447 (2000).
63. Adkins, S., Gan, K. N., Mody, M. & La Du, B. N. Molecular basis for the polymorphic forms of human serum paraoxonase/arylesterase: glutamine or arginine at position 191, for the respective A or B allozymes. *Am. J. Hum. Genet.* **52**, 598–608 (1993).
64. Humbert, R. *et al.* The molecular basis of the human serum paraoxonase activity polymorphism. *Nature Genet.* **3**, 73–76 (1993).
65. Brophy, V. H. *et al.* Effects of 5' regulatory-region polymorphisms on paraoxonase-gene (PON1) expression. *Am. J. Hum. Genet.* **68**, 1428–1436 (2001).
66. Jarvik, G. P. *et al.* Paraoxonase activity, but not haplotype utilizing the linkage disequilibrium structure, predicts vascular disease. *Arterioscler. Thromb. Vasc. Biol.* **23**, 1465–1471 (2003).

Acknowledgements This work was supported by grants from the National Heart, Lung and Blood Institute, the National Institute of Environmental Health Sciences and the National Institute of Mental Health. L.K. is a James S. McDonnell Centennial Fellow. Thanks to D. Altshuler for helpful input, to G. Jarvik, P. Heagerty and P. Scheet for discussions on epistatic risk models, and to T. Banghale, D. Crawford, B. Livingston, R. Mackelprang and M. Rieder for comments on the manuscript.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Predicting disease using genomics

John Bell

University of Oxford, Academic Centre, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK (e-mail: regius@medsci.ox.ac.uk)

Information from the human genome sequence will eventually alter many aspects of clinical practice. It will increase through our understanding of disease mechanisms, and guide the development of new drugs and therapeutic procedures. In the short term, however, knowledge of the genome will have a profound clinical impact on the diagnostic capability of clinical genetics laboratories. Molecular phenotyping using genetic and genomic information will allow early and more accurate prediction and diagnosis of disease and of disease progression. Medicine will become oriented towards disease prevention rather than efforts to cure people at late stages of illness.

Leprosy affects 700,000 people each year and is one of the major health problems in developing countries. *Mycobacterium leprae* attacks the nerves of the hands, feet and face of sufferers, and the lack of feeling in these parts leaves those infected prone to injuries and burns. The longer the disease is left untreated, the more likely it is that deformities will occur. Recently reported genetic research has provided a powerful insight into the cause of this disease, and common genetic risk factors have been identified¹. This information will guide future research aimed at controlling leprosy.

Similar successes in understanding genetic and genomic markers for asthma², cancer^{3–7}, diabetes^{8,9} and cardiovascular disease¹⁰ have been reported. These achievements suggest that new genomic tools will contribute to enhanced clinical practice for many common diseases. Translating this knowledge into routinely applied diagnostics will be a challenge, but will create a new approach to clinical practice with many benefits for patients.

For most diseases, medical treatment doesn't begin until the patient has visited their doctor with various symptoms. By this time it may well be late in the natural history of the disease, and therapeutic intervention may be limited to alleviating symptoms and slowing disease progression (Fig. 1). But in some areas of medicine, this standard approach has progressed to incorporate multiple diagnostic tests that precisely identify disease mechanisms, indicate the most appropriate type of therapeutic intervention, and evaluate therapeutic response and disease outcome. This is particularly evident in the management of infectious diseases such as HIV¹¹ and hepatitis C¹², in which viral load or disease resistance can be rapidly and efficiently measured to help inform therapy, and where diagnostics and therapeutics are closely coupled. In HIV, for example, viral load¹³ is an important indicator of response to therapy or relapse, and mutation screening can help to provide a molecular profile of drug resistance¹¹.

Many non-infectious common diseases may also benefit from a similarly close relationship between diagnostics and therapeutics. New genomic tools emerging from the Human Genome Project will improve our ability to identify individuals at risk or in the presymptomatic phase of diseases, and will more precisely define disease subtypes on the basis of their individual pathophysiology and their responsiveness to therapy. Predictive diagnosis or risk profiling should provide opportunities for environmental modification¹⁴ (such as smoking cessation), early therapy¹⁵ (for example, administering statins for individuals at risk of cardiovascular disease) or targeted cancer screening¹⁶ (for example, the use of colonoscopy in families or individuals at genetic

risk of colorectal cancer). Diagnostic medicine will become increasingly important as our understanding of disease susceptibility and progression markers improves, and as the tools for rapid and effective disease prediction and monitoring are developed.

The human genome has provided a host of opportunities to identify new disease markers. It is already possible to see the impact these tools can have on clinical practice.

Genetics as risk factors

Prediction, prevention and counselling of individuals at risk of genetic diseases have been aimed largely at single-gene disorders that have mendelian patterns of inheritance. These disorders offer the advantage that the identification of genetic variants responsible for disease can lead directly to clinically helpful and reasonably accurate prediction and diagnosis of disease.

Clinical genetics, however, will soon have to move onto aspects of genetics that are less 'deterministic' for a particular disease. Even apparently simple mendelian disorders may prove to have widely variable clinical phenotypes. For example, thalassaemia, an apparently simple genetic disease, has substantial complexities¹⁷. As a result of secondary genetic and environmental factors, individuals with exactly the same globin mutations may suffer either from a severe life-threatening disorder or be relatively unaffected. Another example is when genetic variants identified because of their relatively high penetrance (such as *BRCA1* in breast cancer¹⁸ and *CFTR* in cystic fibrosis¹⁹) have, on closer evaluation, not proved to have consistent levels of penetrance. Even single-gene disorders have a significant level of heterogeneity.

Additional complexity becomes apparent at a molecular level when clinical syndromes are analysed for their genetic causes. A good example of this is the syndrome of sudden cardiac death (SCD), which is often inherited in an autosomal dominant fashion and produces life-threatening cardiac arrhythmias (Fig. 2). Molecular dissection of this phenotype has revealed that multiple mutations of both potassium and sodium channels²⁰, mutations mediating heart muscle disease²¹, or even mutations in proteins anchoring channels to membranes²², can lead to this clinical phenotype. Although comparison across families is difficult, within individual families a single mutation can be tracked, a predictive diagnosis made and appropriate interventions (for example, an implantable cardiac defibrillator or the use of β -blockers to prevent fatal cardiac arrhythmias) can be introduced.

Many common diseases are mechanistically heterogeneous — usually only a small fraction of patients have forms of the disease that are mediated by single genes. For example, the

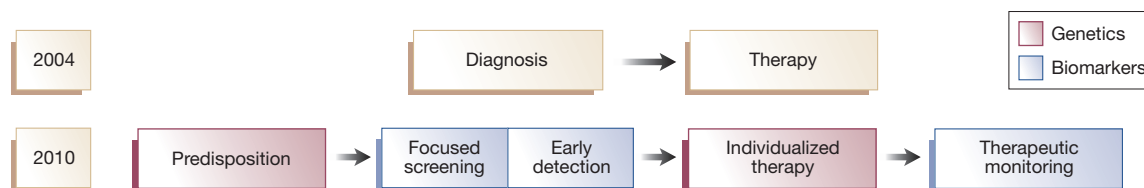


Figure 1 Diagnostic medicine. The role of diagnostics, driven by genetics and the discovery of biomarkers, will grow over the next ten years. Early diagnosis, targeted therapy and disease monitoring will replace the current paradigm of late-stage diagnosis and therapy.

genes for adenomatous polyposis coli (*APC*) and hereditary non-polyposis colorectal cancer (*HNPCC*) account for less than 5% of cases of colorectal cancer; similarly, the diabetes-mediating genes glucokinase (*GCK*), hepatic nuclear factor-1 α (*HNF1A*) and *HNF4A* are responsible for less than 5% of cases of diabetes^{23–26}. Usually, multiple genes and environmental factors mediate disease pathogenesis, and clinical definitions of disease obscure multiple mechanistically distinct subtypes⁸, creating a problem for the application of genetics. Clinical management of these diseases is usually reliant on predicting and managing the risk of disease outcome, the likelihood of a particular disease outcome, or therapeutic efficacy. The evaluation of risk factors defined by physiological, biochemical and radiological parameters is already a critical component of the management of most common diseases. Although the identification and clear validation of a few such risk factors (for example, smoking, hypertension and raised cholesterol levels) and the effect of people acting on that information have had an immense impact on our ability to reduce morbidity and premature death in the population^{9,15}, we still know too few of the risk factors. It is here that genetics may have its greatest impact. Disease-related genetic polymorphisms could be used, in combination with other factors, to define populations and individuals at risk. For example, major susceptibility determinants have been identified in diseases such as asthma², infectious conditions such as leprosy¹, and diabetes^{9,27}. These advantages for patients and practitioners will push clinical genetics towards testing for genetic variants in common diseases, despite their less-than-deterministic contribution to disease.

New approaches to monitoring disease through genomics

Until now, the tools of genetics have been restricted to a clinical setting, but the Human Genome Project has led to the development of substantial new technologies that are capable of defining large sets of biomarkers systematically in biological samples. The systematic approach of analysing all the products of the genome at a messenger RNA or a protein level has emerged from both new detection technology and the availability of the genome sequence^{28,29}. These methodologies, all currently being used in a research setting, will generate data on multiple biomarkers that vary quantitatively very early in disease, with disease onset, with disease progression or with therapeutic response. They may also provide other sets of prognostic factors or biological markers for underlying genetic variants³⁰.

mRNA patterns are strong prognostic markers in cancer, and can be used to define disease subtypes and predict response to therapy^{3–6}. Proteomics appears, even at its existing rudimentary stage, to be able to predict diseases such as cancer⁷, and metabonomics, a systematic approach to analysing small molecules, provides biomarker evidence of vascular disease¹⁰.

Several challenges need to be overcome before these techniques become routine diagnostic tools. Practitioners and regulators are not accustomed to assessing patterns of multiple markers, and so these methods may need to be simplified, as will the technology platforms for routine use. The first chip-based diagnostics are now available for predicting cytochrome P450 metabolism. This enzyme is responsible

for oxidative metabolism of many drugs and is therefore a useful indicator of therapy response. But these techniques will only take off if there is clear evidence of their clinical utility, and if this technology is supported by regulators and practitioners. Ultimately, the education of health care professionals will need to incorporate much more training so that the tools can be used effectively.

Pharmacogenetics

Improving therapeutic efficacy and reducing drug toxicity are two of the most important goals of genomics and genetics in clinical practice. Genetic information may in the future be used widely to more accurately target drugs and to improve their therapeutic usefulness. Consideration of individual variation in metabolism began with Garrod at the beginning of the twentieth century, and a few clear examples (response to suxamethonium, primaquine and debrisoquine) of genetically determined metabolic variation in response to drugs were defined in the 1950s and 1960s (ref. 31). Since then, some metabolic variants have been found to reduce the toxicity of many drugs, not just one (see review in this issue by Evans and Relling, page 464). Genetic variation in cytochrome P450 genes³², acetyltransferase genes, thiopurine methyltransferase³³ and dihydropyrimidine dehydrogenase³⁴ has clinical relevance because in each case it defines patient populations that metabolize drugs at different rates. The availability of genetic and genomic tools will provide important insights into the variation in response and toxicity to many drugs³⁵.

Genetic markers have been used to define some disease subtypes, identify drug targets and predict therapeutic response. For example, characterization of the cytogenetic abnormality that defines chronic myeloid leukaemia has led to the development of a specific drug to combat this disease³⁶ (see review in this issue by Strausberg *et al.*, page 469). The widespread genetic redefinition of disease subtypes is likely to provide many other examples of differential response to therapy. Molecular subtypes of type II diabetes, for example, show differing response to the drug sulphonylureas³⁷, Her2-positive breast cancer is sensitive to the new drug Herceptin³⁸, some molecularly defined subtypes of brain cancer have unusual sensitivity to chemotherapy³⁹, and subtypes of follicular lymphoma defined by transcript profiling are likely to be treatable with the anti-CD20 antibody Rituxan⁴⁰ (see review in this issue by Strausberg *et al.*, page 469).

It may also be possible to predict drug non-responders on the basis of polymorphisms in drug targets or pathways. This approach has been suggested for statin⁴¹, lipo-oxygenase inhibitor⁴² and nicotine-replacement therapy⁴³. Non-metabolic toxicity may be avoided by identifying groups that are particularly vulnerable, because of genetic factors, to complications such as cardiac arrhythmias (long QT genetic variants²⁰), venous thrombosis (factor V Leiden⁴⁴), and potentially fatal systemic drug reactions to abacavir⁴⁵ or carbamazepine³⁵ (HLA). Ultimately, the use of medication to reduce risk (for example, statins could be used to reduce the risk of vascular disease) must be driven by a more complete risk profile, including knowledge of genomic factors (Box 1). Together, these applications of pharmacogenetics will, if established in clinical practice, dramatically alter our use of pharmacological interventions. Few health care

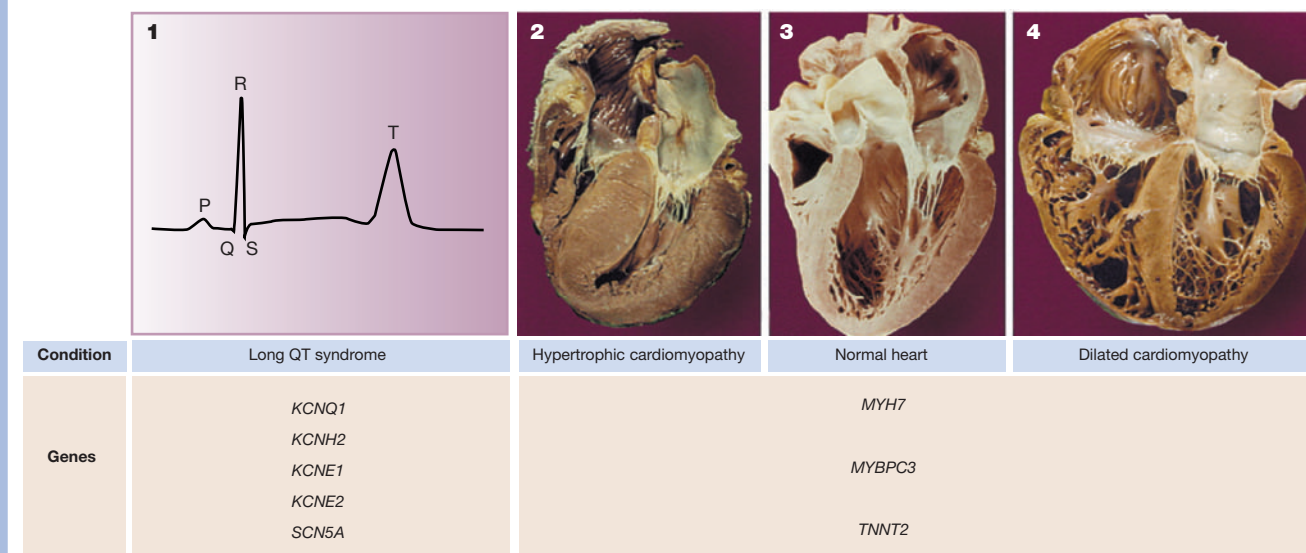


Figure 2 Predictive testing for complex genetic traits in families: sudden cardiac death. SCD arises from three distinct clinical syndromes: long QT syndrome, hypertrophic cardiomyopathy and dilated cardiomyopathy, shown in panels 1, 2 and 4, respectively. Inherited as dominant disorders, individuals at risk can be identified by cascade screening after the causative mutation has been found in a proband.

By screening the limited number of genes related to each of these clinical syndromes, up to 75% of relevant mutations will be detected, and mutation carriers can then be treated with interventions such as β -blockade or implantable cardiac defibrillators to prevent SCD. Panels 2, 3 and 4 reproduced, with permission, from ref. 46.

systems will allow patients to receive drugs that they are unlikely to benefit from, and new information will help to drive genomic diagnostics into routine practice.

From genome to clinical practice

Information valuable to patients and practitioners is emerging from genomic science. This science will need now to be transferred to a clinical setting, creating both technical and cultural challenges for health professionals. Much of the genomic data of clinical relevance generated so far are in a format that is inappropriate for diagnostic testing. The technology has not yet been established for rapid, inexpensive typing of most genomic biomarkers, with the exception of single-nucleotide polymorphisms (SNPs), and too little is known about which SNPs to type for non-mendelian disease. The HapMap, a project to determine common patterns of sequence variation in the genome, is unlikely to contribute to these clinical applications of genetics. This is because varying linkage disequilibrium patterns in different populations make it unsuitable for unsorted patient samples (see review in this issue by Carlson *et al.*, page 446). Simultaneous systematic analysis of larger numbers of biomarkers for disease predic-

tion is still required. These approaches are more suited to research than to routine diagnostic activity. Clear evidence of the predictive strength of these markers in clinical settings remains essential for their implementation.

Very large epidemiological population samples followed prospectively (over a period of years) and characterized for their biomarker and genetic variation will be necessary to demonstrate the clinical utility of these tools. National programmes to develop such genomic epidemiological studies are already in place (Biobank in the United Kingdom) or planned (Canada and China). Obstacles to the routine application of these data in clinical practice include a cultural gap between the approach to clinical practice that is currently employed and that which is possible with these new tools. Diagnostic medicine that includes predisposition testing, early detection, individualized therapy and therapeutic monitoring is neither systematically applied nor well taught in the current health care system. Its implementation will require not just clear data demonstrating its benefits, but also demand by patients and acceptance by health care professionals. This will not come quickly. This approach is also likely to put particular financial pressure on different components of the health care system. The opportunities for clinical genetics to become a mainstream component of clinical medicine are now apparent. This move to the clinic appears to be inevitable, but the transition may take a generation. □

Box 1

Toolbox for clinical genetics/genomics

Genetics/genomics will require the retooling of diagnostic laboratories to incorporate information into disease-prediction or disease-progression diagnostic tests. High-throughput tools will be necessary to test a limited number of markers that have been validated in large clinical studies.

- A DNA sequencing platform
- Comparative genome hybridization (CGH) arrays
- SNP detection technology (for example, dHPLC)
- Multiplex SNP scoring technology
- Transcript profiling technology
- Quantitative multiplex proteomic methodology
- Small-molecule detection assays

doi:10.1038/nature02624

1. Mira, M. T. *et al.* Susceptibility to leprosy is associated with *PARK2* and *PACRG*. *Nature* **427**, 636–640 (2004).
2. Allen, M. *et al.* Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nature Genet.* **35**, 258–263 (2003).
3. van de Vijver, M. J. *et al.* Gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
4. Weigelt, B. *et al.* Gene expression profiles of primary breast tumours maintained in distant metastases. *Proc. Natl Acad. Sci. USA* **100**, 15901–15905 (2003).
5. Wong, Y. F. *et al.* Expression genomics of cervical cancer: Molecular classification and prediction of radiotherapy response by DNA microarray. *Clin. Cancer Res.* **9**, 5486–5492 (2003).
6. Ross, M. E. *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102**, 2951–2959 (2003).
7. Petricoin III, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
8. Bell, G. I. & Polonsky, K. S. Diabetes mellitus and genetically programmed defects in β -cell function. *Nature* **414**, 788–791 (2001).

9. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic β -cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
10. Brindle, J. T. *et al.* Rapid and non-invasive diagnosis of the presence and severity of coronary heart disease using ^1H -NMR-based metabolomics. *Nature Med.* **8**, 1439–1444 (2002).
11. Paul, S. M. & Jorden, V. S. HIV resistance testing. A clinical tool. *N. J. Med.* **100**, 44–49 (2003).
12. Poynard, T., Yeun, M. F., Ratziu, V. & Lai, C. L. Viral hepatitis C. *Lancet* **362**, 2095–2100 (2003).
13. Young, J. *et al.* Stable partnership and progression to AIDS or death in HIV infected patients receiving highly active antiretroviral therapy: Swiss HIV cohort study. *Br. Med. J.* **328**, 15 (2004).
14. The U.S.–Venezuela Collaborative Research Project. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl Acad. Sci. USA* **101**, 3498–3503 (2004).
15. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol-lowering with simvastatin in 5963 people with diabetes: a randomised placebo-controlled trial. *Lancet* **361**, 2005–2016 (2003).
16. Ponz de Leon, M. *et al.* Genetic testing among high-risk individuals in families with hereditary nonpolyposis colorectal cancer. *Br. J. Cancer.* **90**, 882–887 (2004).
17. Weatherall, D. J. Phenotype–genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Rev. Genet.* **2**, 245–255 (2001).
18. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am. J. Hum. Genet.* **62**, 676–689 (1998).
19. Grody, W. W. Molecular genetic risk screening. *Annu. Rev. Med.* **54**, 473–490 (2003).
20. Keating, M. T. & Sanguinetti, M. C. Molecular and cellular mechanisms of cardiac arrhythmias. *Cell* **104**, 569–580 (2001).
21. Watkins, H. Sudden death in hypertrophic cardiomyopathy. *N. Engl. J. Med.* **342**, 422–444 (2000).
22. Mohler, P. J. *et al.* Ankyrin-B mutation causes type 4 long-QT cardiac arrhythmia and sudden cardiac death. *Nature* **421**, 634–639 (2003).
23. Hartge, P. Genes, hormones, and pathways to breast cancer. *N. Engl. J. Med.* **348**, 2352–2354 (2003).
24. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
25. Fajans, S. S., Bell, G. I. & Polonsky, K. S. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N. Engl. J. Med.* **345**, 971–980 (2001).
26. Edlund, H. Transcribing diabetes. *Diabetes* **47**, 1817–1823 (1998).
27. Weedon, M. N. *et al.* Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am. J. Hum. Genet.* **73**, 1208–1212 (2003).
28. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA* **99**, 4465–4470 (2002).
29. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
30. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
31. Weber, W. W. *Pharmacogenetics* (Oxford Univ. Press, New York, 1997).
32. Evans, D. A. P. *et al.* Genetic control of isoniazid metabolism in man. *Br. Med. J.* **2**, 485–491 (1960).
33. Krynetski, E. Y. *et al.* A single point mutation leads to loss of catalytic activity in human thiopurine S-methyltransferase. *Proc. Natl Acad. Sci. USA* **92**, 949–953 (1995).
34. Harris, B. E., Carpenter, J. T. & Diasio, R. B. Severe 5-fluorouracil toxicity secondary to dihydropyrimidine dehydrogenase deficiency. A potential more common pharmacogenetic syndrome. *Cancer* **68**, 499–501 (1991).
35. Chung, W. H. *et al.* Medical genetics: a marker for Stevens–Johnson syndrome. *Nature* **428**, 486 (2004).
36. Savage, D. G. & Antman, K. H. Imatinib mesylate — A new oral targeted therapy. *N. Engl. J. Med.* **346**, 683–693 (2002).
37. Pearson, E. R. *et al.* Genetic cause of hyperglycaemia in response to treatment in diabetes. *Lancet* **362**, 1275–1281 (2003).
38. Fornier, M., Risio, M., Van Poznak, C. & Siedman, A. HER2 testing and correlation with efficacy of trastuzumab therapy. *Oncology* **16**, 1340–1348; 1351–1352; 1355–1358 (2002).
39. Esteller, M. *et al.* Inactivation of the DNA-repair gene *MGMT* and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343**, 1350–1354 (2000).
40. Bohen, S. P. *et al.* Variation in the gene expression patterns in follicular lymphoma and the response to rituximab. *Proc. Natl Acad. Sci. USA* **100**, 1926–1930 (2003).
41. Kuivenhoven, J. A. *et al.* The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. *N. Engl. J. Med.* **338**, 86–93 (1998).
42. Drazen, J. M. *et al.* Pharmacogenetic association between *ALOX5* promoter genotype and the response to anti-asthma treatment. *Nature Genet.* **22**, 168–170 (1999).
43. Johnstone, E. C. *et al.* Genetic variation in dopaminergic pathways and short-term effectiveness of the nicotine patch. *Pharmacogenetics* (in the press).
44. Seligsohn, U. & Lubetsky, A. Genetic susceptibility to venous thrombosis. *N. Engl. J. Med.* **344**, 1222–1231 (2001).
45. Mallan, S. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
46. Seidman, J. G. & Seidman, C. The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell* **104**, 557–567 (2001).

Competing interests statement The author declares competing financial interests: details accompany the paper on www.nature.com/nature.

Epigenetics in human disease and prospects for epigenetic therapy

Gerda Egger, Gangning Liang, Ana Aparicio & Peter A. Jones

Departments of Biochemistry and Molecular Biology and Urology, USC/Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, 1441 Eastlake Avenue, Room 8302L, Los Angeles, California 90089-9181, USA
(e-mail: jones_p@ccnt.hsc.usc.edu)

Epigenetic mechanisms, which involve DNA and histone modifications, result in the heritable silencing of genes without a change in their coding sequence. The study of human disease has focused on genetic mechanisms, but disruption of the balance of epigenetic networks can cause several major pathologies, including cancer, syndromes involving chromosomal instabilities, and mental retardation. The development of new diagnostic tools might reveal other diseases that are caused by epigenetic alterations. Great potential lies in the development of 'epigenetic therapies' — several inhibitors of enzymes controlling epigenetic modifications, specifically DNA methyltransferases and histone deacetylases, have shown promising anti-tumorigenic effects for some malignancies.

The term 'epigenetics' defines all meiotically and mitotically heritable changes in gene expression that are not coded in the DNA sequence itself. Three systems, including DNA methylation, RNA-associated silencing and histone modification, are used to initiate and sustain epigenetic silencing. Unravelling the relationships between these components has led to surprising and rapidly evolving new concepts, showing how they interact and stabilize each other (Fig. 1). Disruption of one or other of these interacting systems can lead to inappropriate expression or silencing of genes, resulting in 'epigenetic diseases'. Here we discuss potential causes of some of these diseases, and suggest how they might be treated in the future.

Methylation of the C⁵ position of cytosine residues in DNA has long been recognized as an epigenetic silencing mechanism of fundamental importance^{1,2}. The methylation of CpG sites within the human genome is maintained

by a number of DNA methyltransferases (DNMTs) and has multifaceted roles for the silencing of transposable elements, for defence against viral sequences and for the transcriptional repression of certain genes. 5-Methylcytosine is highly mutagenic, causing C:G to T:A transitions and resulting in a strong suppression of the CpG methyl-acceptor site in human DNA (Box 1). CpG islands, which are regions of more than 500 base pairs in size and with a GC content greater than 55% (ref. 3), have been conserved during evolution because they are normally kept free of methylation. These stretches of DNA are located within the promoter regions of about 40% of mammalian genes and, when methylated, cause stable heritable transcriptional silencing. Aberrant *de novo* methylation of CpG islands is a hallmark of human cancers and is found early during carcinogenesis⁴.

Histone modifications have also been defined as epigenetic modifiers. Post-translational modifications of histones, including acetylation and methylation of conserved lysine residues on the amino-terminal tail domains, have been studied closely over the past few years. Generally, the acetylation of histones marks active, transcriptionally competent regions, whereas hypoacetylated histones are found in transcriptionally inactive euchromatic or heterochromatic regions. Histone methylation can be a marker for both active and inactive regions of chromatin. Methylation of lysine 9 on the N terminus of histone H3 (H3-K9) is a hallmark of silent DNA and is globally distributed throughout heterochromatic regions such as centromeres and telomeres. It is also found on the inactive X chromosome and at silenced promoters⁵. In contrast, methylation of lysine 4 of histone H3 (H3-K4) denotes active and is found predominantly at promoters of active genes⁵. Because lysine methylation can be monomeric, dimeric or trimeric, and histones may also be subject to other post-translational modifications such as phosphorylation⁶, this enormous variation leads to a multiplicity of possible combinations of different modifications. This might constitute a 'histone code'⁷, which can be read and interpreted by different cellular factors.

Links between histone modifications and DNA methylation have been found in plants and fungi (Fig. 1), where H3-K9 methylation is a prerequisite for DNA methylation^{8–10}. DNA methylation can also trigger H3-K9 methylation^{11–13}, and this has also been shown in mammals,

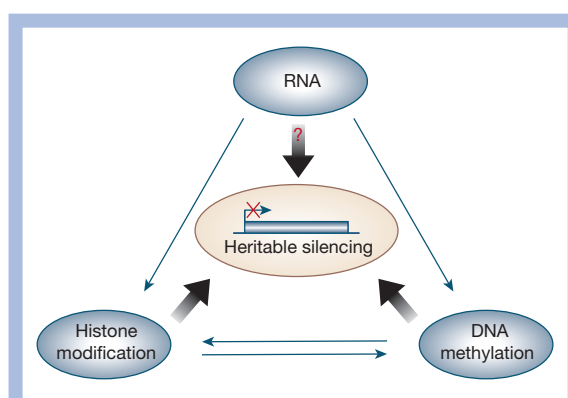
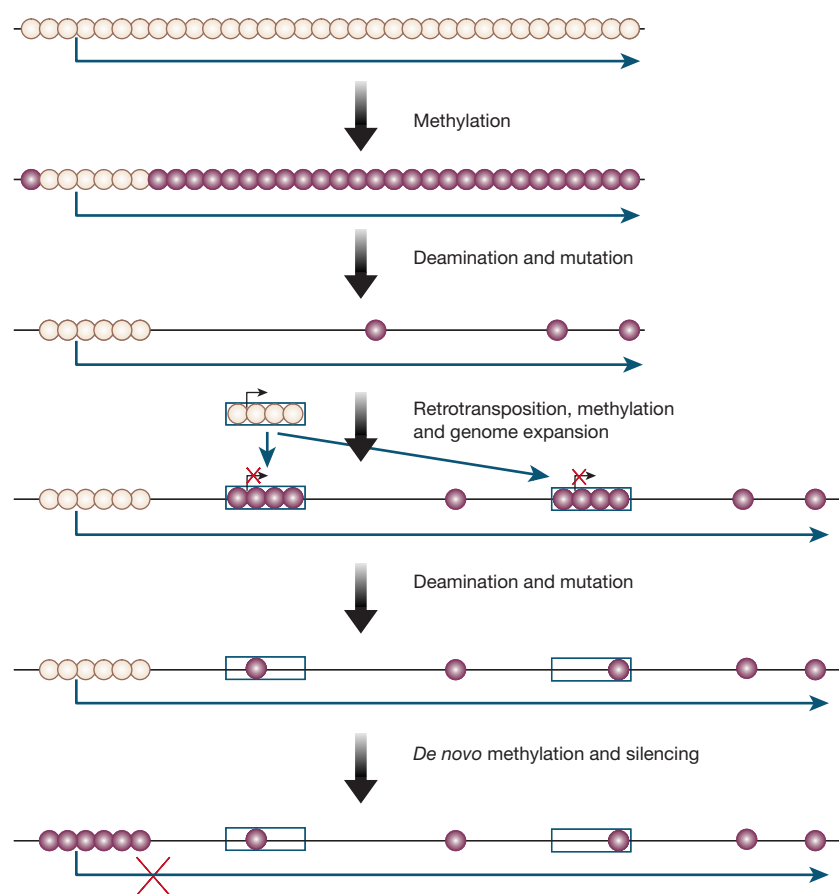


Figure 1 Interaction between RNA, histone modification and DNA methylation in heritable silencing. Histone deacetylation and other modifications, particularly the methylation of lysine 9 within histone H3 (H3-K9) residues located in the histone tails, cause chromatin condensation and block transcriptional initiation. Histone modification can also attract DNA methyltransferases to initiate cytosine methylation, which in turn can reinforce histone modification patterns conducive to silencing. Experiments in yeast and plants have clearly shown the involvement of RNA interference in the establishment of heterochromatic states and silencing. RNA triggering of heritable quiescence might therefore also be involved in higher organisms.

Box 1

Evolution and methylation of CpG islands

Early genomes did not contain 5-methylcytosine, and CpG sites (cream circles in the figure) occurred as frequently as expected on a statistical basis. As genomes become methylated at CpG sites (pink circles), promoters of about half of human genes were somehow protected from this modification in the germ line and remained as CpG islands. Because of the well-known enhanced mutability of methylated cytosine, CpGs were converted to TpGs and depleted from the rest of the genome, particularly in the transcribed regions of genes in which most of the methylation occurs. About 40% of the human genome is made up of transposable elements (four cream circles in a group) or their relics, and the active elements are capable of inserting themselves both into genes and into heterochromatic regions. Methylation of the CpG sites within these transposed elements results in the silencing of their promoters at the same time as it does not hinder the transcription of the host gene. In this way, mammals are different from *Neurospora crassa*, in which methylation of cytosine residues in the transcribed regions blocks elongation by polymerase II. Organisms such as *Drosophila melanogaster*, which have very little cytosine methylation, show a higher degree of transposition of transposable elements⁹¹. The methylation of transposable elements such as Alus, LINEs and LTR sequences increases the rate of C to T transition mutations at these sites, so these sequences eventually become depleted of CpGs. The interaction between transposable elements and the cytosine methylation system results in genomic expansion and CpG depletion. Physiological methylation of CpG islands in the



promoters of X-linked or imprinted genes, for example, leads to their long-term silencing. Pathological methylation can result in the silencing of tumour-suppressor or other cancer-relevant genes. Epigenetic therapy seeks to reverse these changes.

where H3-K9 methylation directs DNA methylation to pericentromeric heterochromatin¹⁴. Interactions between histone deacetylases (HDACs), histone methyltransferases and methylcytosine-binding proteins^{15,16} lead to the recruitment of DNA methyltransferases¹⁷, although it is not yet clear what initiates the recruitment of the different epigenetic modifiers to their specific target sequences.

The role of RNA in post-transcriptional silencing has attracted much interest. However, RNA, in the form of antisense transcripts, noncoding RNAs (such as *Xist*) or RNA interference (RNAi), can also lead to mitotically heritable transcriptional silencing by the formation of heterochromatin. In *Schizosaccharomyces pombe*, for example, the deletion of different components of the RNAi machinery results in an impairment of centromere function, a derepression of transgenes integrated at centromeres, and a loss of the characteristic H3-K9 methylation within the same region^{18,19}. A similar connection between RNAi, DNA methylation and H3-K9 methylation has been demonstrated in *Arabidopsis thaliana*²⁰. Although RNAi-directed silencing of heterochromatic regions has not yet been shown in mammals, the involvement of RNA in different silencing mechanisms has been described. For example, antisense RNAs are involved in the silencing of some mammalian imprinted genes²¹ and in dosage compensation in mammals²². A recent report of

a case of α -thalassaemia showed how antisense transcription could lead to DNA methylation and stable silencing of a globin gene²³. RNA might therefore be a key trigger to direct histone modifications (for example, H3-K9 methylation) and DNA methylation to specific loci (for example, pericentromeric heterochromatin), thereby evoking heritable and stable silencing. The therapeutic activation of abnormally silenced genes thus requires drugs that can target the inherent stability of these multifaceted changes.

Epigenetic diseases

Multicellular organisms require mutually reinforcing mechanisms permitting heritable patterns of gene silencing. Mutations in genes that affect global epigenetic profiles can give rise to human diseases, which can be inherited or somatically acquired (Table 1). Interestingly, many of these epigenetic abnormalities result in chromosomal alterations and learning disabilities. For example, mutations in the *ATRX* gene result in consistent changes in the pattern of methylation of ribosomal DNA, Y-specific repeats and subtelomeric repeats. Fragile X syndrome results when a CGG repeat in the *FMR1* 5' untranslated region expands and becomes methylated *de novo*, causing the gene to be silenced and creating a visible 'fragile' site on the X chromosome under certain conditions. On a more global level, the ICF (immunodeficiency, centromeric region instability and facial anomalies)

Table 1 Epigenetic diseases

Disease	Symptom	Aetiology	References
ATR-X syndrome	Intellectual disabilities, α -thalassaemia	Mutations in <i>ATRX</i> gene, hypomethylation of certain repeat and satellite sequences	82
Fragile X syndrome	Chromosome instability, intellectual disabilities	Expansion and methylation of CGG repeat in <i>FMR1</i> 5' UTR, promoter methylation	83
ICF syndrome	Chromosome instability, immunodeficiency	<i>DNMT3b</i> mutations, DNA hypomethylation	84
Angelman's syndrome	Intellectual disabilities	Deregulation of one or more imprinted genes at 15q11–13 (maternal)	85
Prader-Willi syndrome	Obesity, intellectual disabilities	Deregulation of one or more imprinted genes at 15q11–13 (paternal)	86
BWS	Organ overgrowth	Deregulation of one or more imprinted genes at 11p15.5 (e.g. <i>IGF2</i>)	87
Rett syndrome	Intellectual disabilities	<i>MeCP2</i> mutations	25, 26
α -Thalassaemia (one case)	Anaemia	Methylation of α 2-globin CpG island, deletion of <i>HBA1</i> and <i>HBQ1</i>	23
Various cancers	Microsatellite instability	<i>De novo</i> methylation of <i>MLH1</i>	29
	Disruption of Rb, p53 pathway, uncontrolled proliferation	<i>De novo</i> methylation of various gene promoters	4
	Disruption of SWI-SNF chromatin remodelling complex	Mutations in <i>SNF5</i> , <i>BRG1</i> , <i>BRM</i>	36
	Overexpression of <i>IGF2</i> , silencing of <i>CDKN1C</i>	Loss of imprinting	88, 89
Leukaemia	Disturbed haematopoiesis	Chromosomal translocations involving HATs and HMTs	62
Rubinstein-Taybi syndrome	Intellectual disabilities	Mutation in CREB-binding protein (histone acetylation)	90
Coffin-Lowry syndrome	Intellectual disabilities	Mutation in <i>Rsk-2</i> (histone phosphorylation)	90

ATR-X syndrome, α -thalassaemia, mental retardation syndrome, X linked; BWS, Beckwith-Wiedemann syndrome; CREB, cAMP-response-element-binding protein; HAT, histone acetyltransferase; HMT, histone methyltransferase; ICF, immunodeficiency, centromeric region instability and facial anomalies syndrome; UTR, untranslated region.

syndrome is caused by mutations in the *DNMT3b* gene, which is an essential enzyme required for the establishment of DNA methylation patterns²⁴. The fact that all of these diseases show gross chromosomal anomalies points to a central role for epigenetic mechanisms in chromosome architecture.

Several inherited syndromes are due to faulty genomic imprinting — defined as parent-specific, monoallelic expression of a gene — such as Angelman's syndrome, Prader-Willi syndrome and Beckwith-Wiedemann syndrome (BWS). In these conditions, an abnormal phenotype is established as a result of the absence of the paternal or maternal copy of an imprinted gene or because of deregulation of an imprinted gene.

For example, a cluster of imprinted genes at 11p15.5 is involved in the pathology of BWS, a syndrome characterized by organ overgrowth and association with embryonal tumours such as Wilms' tumour. Loss of methylation in imprinting control regions can cause a deregulation of imprinting and either biallelic expression (such as *IGF2*) or silencing (such as *CDKN1C*) of imprinted genes, which is found in most sporadic cases of BWS.

Particular interest has recently been generated by the finding that one of the most common forms of intellectual disability in young girls, namely Rett syndrome, is due to germline mutations of *MeCP2* (ref. 25). The *MeCP2* protein binds to methylcytosine residues²⁶ and the disease pathogenesis might be linked to the derepression of genes normally suppressed by DNA methylation. No direct evidence has yet been found for this, because human cells with *MeCP2* mutations do not show global gene derepression²⁶. However, *MeCP2* might have a key role in the control of neuronal gene activity resulting in the pathology of Rett syndrome^{27,28}.

As mentioned above, a fascinating recent discovery was that the transcription of an antisense RNA led to gene silencing and to the methylation of a structurally normal α -globin gene in a patient with thalassaemia²³. Other examples of inappropriate gene silencing might contribute to disease yet might be missed as part of conventional diagnosis. This case of thalassaemia might well be the tip of the iceberg and indicates that several other kinds of human disease might be the result of genomic rearrangements that cause epigenetic silencing.

Epigenetic changes can also have a major role in the development of human cancer. For example, a high percentage of patients with sporadic colorectal cancers with a microsatellite instability phenotype show methylation and silencing of the gene encoding *MLH1* (ref. 29). Thus, epigenetic silencing can result directly in genetic instability. In some cases, promoter-associated methylation of *MLH1* is found not only in tumour but also in normal somatic tissues, including spermatozoa. These germline 'epimutations' predispose individuals carrying aberrant

methylation patterns to multiple cancers^{30,31}. Disruption of pathways that lead to cancer is often caused by the *de novo* methylation of the relevant gene's promoters⁴. Epigenetic silencing has been recognized as a third pathway satisfying Knudson's hypothesis that two hits are necessary for the silencing of tumour-suppressor genes³².

Chromatin-modifying enzymes have also been associated with the aetiology of different haemopathologies. A characteristic of human leukaemias is the presence of various chromosomal translocations, leading to the expression of fusion proteins. Both histone acetyltransferases and histone methyltransferases can be part of such fusions and cause the upregulation of target genes³³. In acute promyelocytic leukaemia, the oncogenic fusion protein PML-RAR α (promyelocytic leukaemia-retinoic acid receptor- α) recruits an HDAC to repress genes essential for the differentiation of haematopoietic cells³⁴. Similarly, in acute myeloid leukaemia, AML1-ETO fusions recruit the repressive N-CoR-Sin3-HDAC1 complex and inhibit myeloid development³⁵.

The importance of correct chromatin composition is further underlined by the roles of ATP-dependent chromatin remodelling complexes in disease. These are multisubunit complexes that are capable of moving and shifting nucleosomes, thereby regulating transcription. Several members of the highly conserved SWI-SNF complex have been implicated in cancer³⁶. For example, a loss of *SNF5* is observed in paediatric cancers, and the ATPase subunits *BRM* and *BRG1* are mutated in a variety of cancer cell lines and primary tumours; this is associated with a poorer prognosis in patients with non-small-cell lung cancer³⁶.

Epigenetic drift

A distinguishing feature of epigenetic changes in comparison with genetic changes is that they tend to be acquired in a gradual rather than an abrupt process. For example, a generalized decrease in genomic 5-methylcytosine concentrations occurs as mammalian cells age^{37,38}, and this gradual loss of DNA methylation can result in aberrant gene activation. The decrease in 5-methylcytosine occurs at the same time as there is a localized hypermethylation of CpG islands at gene promoters³⁹. The accumulation of methylated CpGs within the promoter might be acting as a 'rheostat' rather than a 'switch' for gene silencing⁴⁰. These alterations are therefore targets for prevention strategies. It has already been shown that a lowering of DNA methylation or DNA methyl-binding proteins can decrease the number of intestinal polyps in cancer-prone mice⁴¹. Although the full extent of epigenetic changes is largely unknown, there are obvious advantages to designing strategies that can prevent the aberrant regulation of genes as a function of age.

Table 2 Epigenetic drugs

Target	Drug	Clinical trials
DNA methylation	5-Azacytidine	Phase I/II/III
	5-Aza-2'-deoxycytidine	Phase I/II/III
	FCDR	
	Zebularine	
	Procainamide	
	EGCG	Phase I
Histone deacetylase	Psammaplin A	
	Antisense oligomers	Phase I
	Many ⁶⁵ , including:	
	Phenylbutyric acid	Phase I/II
	SAHA	Phase I/II
	Depsipeptide	Phase I/II
	Valproic acid	Phase I/II

EGCG, epigallocatechin-3-gallate; FCDR, 5-fluoro-2'-deoxycytidine; SAHA, suberoylanilide hydroxamic acid.

Epigenetic therapy

The fact that many human diseases, including cancer, have an epigenetic aetiology has encouraged the development of a new therapeutic option that might be termed 'epigenetic therapy.' Many agents have been discovered that alter methylation patterns on DNA or the modification of histones, and several of these agents are currently being tested in clinical trials (Table 2).

Inhibitors of DNA methylation rapidly reactivate the expression of genes that have undergone epigenetic silencing, particularly if this silencing has occurred in a pathological situation. The prototype inhibitors, 5-azacytidine (5-aza-CR) and 5-aza-2'-deoxycytidine (5-aza-CdR), were initially developed as cytotoxic agents⁴², but it was subsequently discovered that they are powerful inhibitors of DNA methylation and induce gene expression and differentiation in cultured cells^{43,44}. Both nucleoside analogues are converted to the deoxynucleotide triphosphates and are then incorporated in place of cytosine into replicating DNA (Fig. 2). They are therefore active only in S-phase cells⁴³, where they serve as powerful mechanism-based inhibitors of DNA methylation⁴⁴. DNA methyltransferases get trapped on DNA containing modified bases such as azacytosine, 5-fluorocytosine, pseudoisocytosine or zebularine, resulting in the formation of heritably demethylated DNA^{44,45}. Covalent attachment of the various DNA methyltransferases to DNA might well be responsible for the cytotoxicities of these agents, particularly at high doses⁴⁶.

Initial clinical trials used increasingly cytotoxic doses of these agents, even though the induction of gene expression shows a bell-shaped response curve⁴⁴. Exciting new clinical trials have shown that low doses of these agents might be efficacious in treating myeloid dysplastic syndrome and other leukaemias⁴⁷. In addition, azanucleosides are being tested for the treatment of haemoglobinopathies⁴⁸, in which the aim is to cause the demethylation of the promoters of the fetal globin genes that have become silenced in the patient as part of normal development, leading to an increase in haemoglobin F to correct the anaemia that characterizes these diseases.

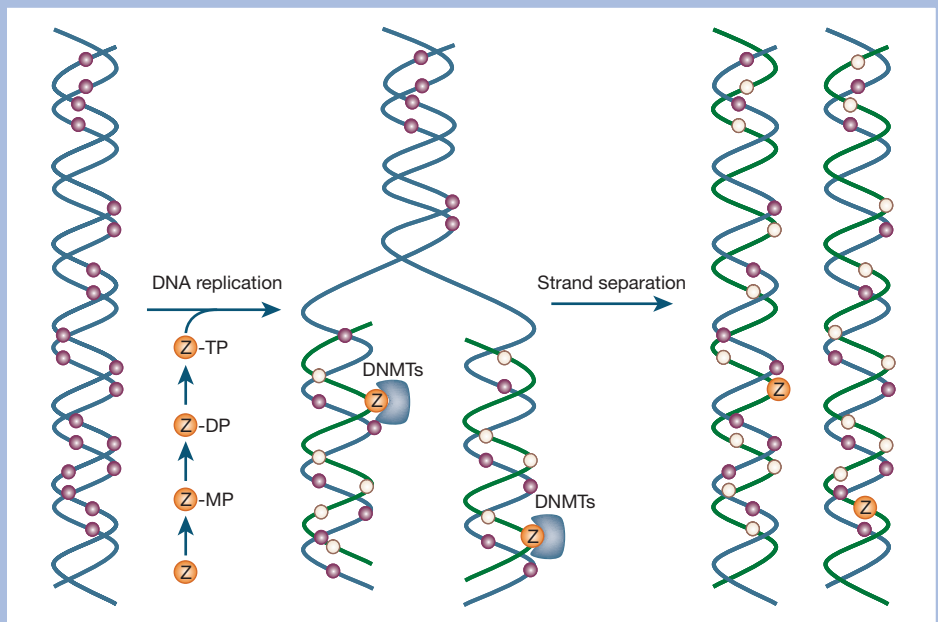
A disadvantage of the azanucleosides is their instability in aqueous solutions, but this might be overcome by the use of other analogues, such as zebularine or 5-fluoro-2'-deoxycytidine, which also inhibit DNA methylation after incorporation into DNA and might be orally active⁴⁹. Procainamide, which is used to treat cardiac arrhythmias, is also an inhibitor of DNA methylation⁵⁰. In addition, natural products derived from tea⁵¹ and from sponges⁵² have shown activity *in vitro*. Clinical trials with antisense oligonucleotides that target the DNA methyltransferases are also underway⁵³.

Epigenetic silencing is almost universally associated with histone deacetylation, which is catalysed by at least three classes of HDACs in human cells. The HDACs are partly redundant in function, and a growing series of small molecules has been designed to inhibit their activities either globally or more specifically (Table 2). HDAC inhibitors can induce differentiation, growth arrest and/or apoptosis in transformed cells in culture and in tumours. The driving hypothesis is that accumulation of acetylated proteins, particularly histones, results in the induction of genes and the upregulation of others that have become epigenetically silenced. In particular, the gene encoding p21, which is a cell-cycle kinase inhibitor, is commonly upregulated in tumour cells treated with these agents in the absence of p53 (ref. 54). This is important in terms of cancer therapy, because many cancers have no functional p53 and are therefore unable to arrest cells in a p53-dependent fashion. Different HDAC inhibitors are being used intravenously or orally in several phase I and II clinical trials⁵⁵, in which changes in histone acetylation have been documented. Because there are many different HDACs, it will be important in the future to design therapies that can target individual enzymes and thus increase the precision of this approach.

Coupling therapies

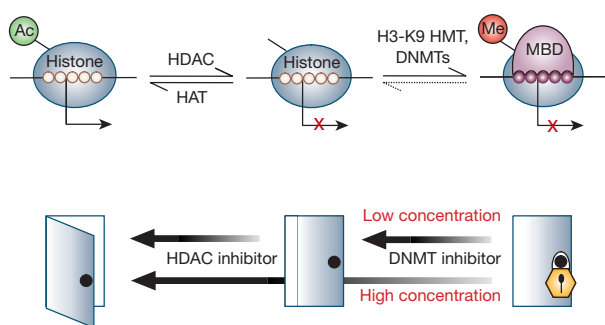
The links between histone modification and DNA methylation (see Fig. 1) have encouraged investigators to think about dual therapies

Figure 2 Mechanism of action of nucleoside analogue inhibitors. Deoxynucleoside analogues such as 5-aza-2'-deoxycytidine (depicted by Z) are converted into the triphosphate inside S-phase cells and are incorporated in place of cytosine into DNA. Ribonucleosides such as 5-azacytidine or zebularine are reduced at the diphosphate level by ribonucleotide reductase for incorporation (not shown). Once in DNA, the fraudulent bases form covalent bonds with DNA methyltransferases (DNMTs), resulting in the depletion of active enzymes and the demethylation of DNA. Pink circles, methylated CpG; cream circles, unmethylated CpG.



Box 2

Gene silencing and pharmacological reactivation



Genes containing CpG islands in their promoters (for example, the p16 tumour-suppressor gene) can be reversibly controlled (opened or shut) by altered levels of histone acetylation and the presence or absence of transcription factors at the promoter. Heritable silencing (locking) is achieved by multiple histone modification changes, including trimethylation of the H3-K9, binding of methylated DNA-binding proteins such as MeCP2, and *de novo* methylation of the CpG island⁹². The exact order in which these changes are acquired is not certain, although it seems likely that H3-K9 methylation precedes cytosine methylation⁹³. Once these changes have occurred, they tend to reinforce each other and the gene becomes refractory to reactivation (upper panel). Demethylated, silenced genes can often be activated by histone deacetylase (HDAC) inhibitors, but these agents are largely ineffective with methylated promoters⁵⁸. In contrast, DNA methylation inhibitors are highly effective not only at removing the cytosine methylation but also at rapidly reversing the chromatin structural changes^{94,95}, and not only in unlocking the gene but also in opening it for transcription. New clinical trials seek to take advantage of the synergy between low doses of DNA methylation inhibitors and HDAC inhibitors (lower panel). DNMT, DNA methyltransferase; HAT, histone acetyltransferase; HMT, histone methyltransferase; MBD, methyl-CpG-binding domain protein.

combining DNA methylation inhibitors with HDAC inhibitors (Box 2). Jahangeer *et al.*⁵⁶ first showed that 5-aza-CR and butyrate (an HDAC inhibitor) acted synergistically to upregulate the β -adrenergic receptor in HeLa cells. Subsequently, Ginder *et al.*⁵⁷ showed that the two agents acted synergistically in anaemic chickens to increase the level of embryonic p-type globin messenger RNA in the haematopoietic cells. The synergy of demethylation and HDAC inhibitors was investigated in greater detail by Cameron *et al.*⁵⁸. Suzuki *et al.*⁵⁹ and Yamashita *et al.*⁶⁰ have used the approach of treating cultured cells with 5-aza-CdR and trichostatin A to isolate new tumour-suppressor genes efficiently. In addition, a recent study has shown a strong synergy between 5-aza-CdR and phenyl butyrate (an HDAC inhibitor) in the prevention of murine lung cancer⁶¹. This is therefore an exciting time, as clinical trials are developed that can target these two epigenetic modifications in patients⁶². In addition to these approaches, it might be advantageous to sensitize cells by epigenetic therapy followed by treatment with chemotherapy⁶³, interferon⁶⁴ or immunotherapy⁶⁵, among others. In terms of cancer therapy, it will be essential to include both genetic (see reviews in this issue by Strausberg *et al.* (page 469) and Bell (page 453)) and epigenetic markers⁶⁶, which will permit an individually targeted therapy.

Potential pitfalls of epigenetic therapy

Despite the promise of epigenetic therapy, there are several concerns regarding the clinical applications of these agents. These relate mainly to the nonspecific activation of genes and transposable elements in

normal cells, and also to potential mutagenicity and carcinogenicity. Unfortunately, few studies have examined the effects of azanucleosides on completely normal cells as opposed to cell lines. The drugs have profound effects on immortal lines⁴⁴, yet only 0.4% of 6,600 genes (compared with 1% in tumour cell lines) analysed were upregulated more than fourfold in normal human fibroblasts exposed to 5-aza-CdR⁶⁷. Early studies showed that 5-aza-CR could activate a human X chromosome in a rodent somatic cell hybrid⁶⁸ but not in normal human cells⁶⁹. These data suggest that DNA methylation is only one of the mechanisms enforcing silencing in normal cells (Fig. 1) and that they are therefore less sensitive to drug-induced gene activation. Imprinted genes can be activated by 5-aza-CdR⁷⁰, implying a need for caution. Azanucleosides have been shown to be mutagenic⁷¹ and possibly carcinogenic in rats⁷², and might be able to activate silenced oncogenes⁷³, although they can clearly act as cancer-prevention agents^{41,61}.

Azanucleosides have shown some promising results in clinical trials without much evidence of adverse effects. For example, treatment of 41 leukaemia patients with 5-aza-CdR showed only mild effects on global genomic demethylation, as measured by changes in Alu methylation⁷⁴. Original methylation levels were regained within two weeks after therapy, and no development of a secondary malignancy was observed in follow-up studies. Furthermore, administration of a low dose of 5-aza-CdR induced cytogenetic remissions in a substantial number of patients with myelodysplastic syndrome with pre-existing chromosomal abnormalities⁷⁵. No increase in chromosomal instability was observed after therapy, which argues against a strong effect of 5-aza-CdR on chromosomal integrity in patients.

The therapeutic mechanism of action of DNA methylation and HDAC inhibitors is by no means straightforward. Evidently, both classes of agents can activate genes — but is this the way that they work in patients? HDAC inhibitors and DNA methylation inhibitors are cytotoxic agents and induce cell-cycle arrest and apoptosis by upregulating p21 and/or p53 (refs 55, 76). Loss of genomic methylation causes p53-dependent apoptosis⁷⁷, and p53 represses DNMT1 (ref. 78), suggesting a feedback loop between the two proteins. In addition, the binding of proteins, including DNA methyltransferases, to the DNA of cells treated with azanucleosides can result in cytotoxicity^{46,79}. These points make it imperative that surrogate endpoints be examined in patients to gain a better understanding of the mechanism of action.

Future perspectives

Epigenetic disorders give rise to several significant human diseases, and the race is on to find therapies that can reverse silencing. Sometimes it might be possible to reuse a wild-type embryonic gene in place of a mutated adult gene such as in thalassaemia or sickle-cell anaemia. Rett syndrome is caused by mutations in the *MeCP2* gene, which resides on the X chromosome. Because X-inactivation is mostly random, about 50% of the cells in these girls harbour a suppressed wild-type gene, so a valid target might be the reactivation of this good but unused copy. Because of the potential risks of epigenetic therapy, it is likely that trials to validate the approach will be based on patients with life-threatening diseases such as cancer. Here, it is important to remember that the target is an abnormally methylated CpG island and that these seem to be particularly sensitive to reactivation by DNA methylation inhibitors in cancer cells. In addition, because multiple genes become methylated in individual cancers⁸⁰, there is the possibility of hitting many targets with one drug. Finally, because methylation of CpG islands increases with age³⁹ and could therefore contribute to the development of chronic diseases in addition to cancer, there might be benefits in drugs and lifestyle changes that could bring about reversion, or slow gradual epigenetic silencing.

It is apparent that we are just at the beginning of understanding the substantial contributions of epigenetics to human disease, and there are probably many surprises ahead. For example, the finding

that loss of imprinting can be seen not only in normal colonic epithelium but also in the lymphocytes of colorectal patients was completely unexpected⁸¹. Elucidating the whole bandwidth of epigenetic mechanisms is an exciting challenge and will eventually lead to a clearer understanding of the development of human disease and direct therapeutic concepts into new directions.

doi:10.1038/nature02625

- Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).
- Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25 (1975).
- Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA* **99**, 3740–3745 (2002).
- Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Rev. Genet.* **3**, 415–428 (2002).
- Lachner, M. & Jenuwein, T. The many faces of histone lysine methylation. *Curr. Opin. Cell Biol.* **14**, 286–298 (2002).
- Fischle, W., Wang, Y. & Allis, C. D. Binary switches and modification cassettes in histone biology and beyond. *Nature* **425**, 475–479 (2003).
- Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
- Tamaru, H. & Selker, E. U. A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature* **414**, 277–283 (2001).
- Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**, 556–560 (2002).
- Malagnac, F., Bartee, L. & Bender, J. An *Arabidopsis* SET domain protein required for maintenance but not establishment of DNA methylation. *EMBO J.* **21**, 6842–6852 (2002).
- Johnson, L., Cao, X. & Jacobsen, S. Interplay between two epigenetic marks. DNA methylation and histone H3 lysine 9 methylation. *Curr. Biol.* **12**, 1360–1367 (2002).
- Soppe, W. J. et al. DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J.* **21**, 6549–6559 (2002).
- Tariq, M. et al. Erasure of CpG methylation in *Arabidopsis* alters patterns of histone H3 methylation in heterochromatin. *Proc. Natl Acad. Sci. USA* **100**, 8823–8827 (2003).
- Lehnertz, B. et al. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.* **13**, 1192–1200 (2003).
- Nan, X. et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).
- Fuks, F. et al. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem.* **278**, 4035–4040 (2003).
- Fuks, F., Burgers, W. A., Brehm, A., Hughes-Davies, L. & Kouzarides, T. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nature Genet.* **24**, 88–91 (2000).
- Hall, I. M. et al. Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237 (2002).
- Volpe, T. A. et al. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
- Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299**, 716–719 (2003).
- Rougeulle, C. & Heard, E. Antisense RNA in imprinting: spreading silence through Air. *Trends Genet.* **18**, 434–437 (2002).
- Panning, B. & Jaenisch, R. RNA and the epigenetic regulation of X chromosome inactivation. *Cell* **93**, 305–308 (1998).
- Tufarelli, C. et al. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genet.* **203**, 157–165 (2003).
- Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99**, 247–257 (1999).
- Amir, R. E. et al. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nature Genet.* **23**, 185–188 (1999).
- Klose, R. & Bird, A. Molecular biology. MeCP2 repression goes nonglobal. *Science* **302**, 793–795 (2003).
- Chen, W. G. et al. Derepression of BDNF transcription involves calcium-dependent phosphorylation of MeCP2. *Science* **302**, 885–889 (2003).
- Martinowich, K. et al. DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. *Science* **302**, 890–893 (2003).
- Kane, M. F. et al. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Res.* **57**, 808–811 (1997).
- Gazzoli, I., Loda, M., Garber, J., Syngal, S. & Kolodner, R. D. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the *MLH1* gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. *Cancer Res.* **62**, 3925–3928 (2002).
- Suter, C. M., Martin, D. I. & Ward, R. L. Germline epimutation of *MLH1* in individuals with multiple cancers. *Nature Genet.* advance online publication 4 April 2004 (doi:10.1038/ng1342).
- Jones, P. A. & Laird, P. W. Cancer epigenetics comes of age. *Nature Genet.* **21**, 163–167 (1999).
- Hake, S. B., Xiao, A. & Allis, C. D. Linking the epigenetic 'language' of covalent histone modifications to cancer. *Br. J. Cancer* **90**, 761–769 (2004).
- Grignani, F. et al. Fusion proteins of the retinoic acid receptor- α recruit histone deacetylase in promyelocytic leukaemia. *Nature* **391**, 815–818 (1998).
- Jones, L. K. & Saba, V. Chromatin modification, leukaemia and implications for therapy. *Br. J. Haematol.* **118**, 714–727 (2002).
- Roberts, C. W. & Orkin, S. H. The SWI/SNF complex—chromatin and cancer. *Nature Rev. Cancer* **4**, 133–142 (2004).
- Wilson, V. L. & Jones, P. A. DNA methylation decreases in aging but not in immortal cells. *Science* **220**, 1055–1057 (1983).
- Richardson, B. C. Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. *J. Nutr.* **132**, 2401S–2405S (2002).
- Issa, J. P. CpG-island methylation in aging and cancer. *Curr. Top. Microbiol. Immunol.* **249**, 101–118 (2000).
- Beaudet, A. L. & Jiang, Y. H. A rheostat model for a rapid and reversible form of imprinting-dependent evolution. *Am. J. Hum. Genet.* **70**, 1389–1397 (2002).
- Laird, P. W. et al. Suppression of intestinal neoplasia by DNA hypomethylation. *Cell* **81**, 197–205 (1995).
- Sorm, F., Piskala, A., Cihak, A. & Vesely, J. 5-Azacytidine, a new, highly effective cancerostatic. *Experientia* **20**, 202–203 (1964).
- Constantinides, P. G., Jones, P. A. & Gevers, W. Functional striated muscle cells from non-myoblast precursors following 5-azacytidine treatment. *Nature* **267**, 364–366 (1977).
- Jones, P. A. & Taylor, S. M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85–93 (1980).
- Zhou, L. et al. Zebularine: a novel DNA methylation inhibitor that forms a covalent complex with DNA methyltransferases. *J. Mol. Biol.* **321**, 591–599 (2002).
- Michalowsky, L. A. & Jones, P. A. Differential nuclear protein binding to 5-azacytosine-containing DNA as a potential mechanism for 5-aza-2'-deoxycytidine resistance. *Mol. Cell. Biol.* **7**, 3076–3083 (1987).
- Issa, J. P. et al. Phase I study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2'-deoxycytidine (decitabine) in hematopoietic malignancies. *Blood* **103**, 1635–1640 (2004).
- Sauntharajah, Y. et al. Effects of 5-aza-2'-deoxycytidine on fetal hemoglobin levels, red cell adhesion, and hematopoietic differentiation in patients with sickle cell disease. *Blood* **102**, 3865–3870 (2003).
- Cheng, J. C. et al. Inhibition of DNA methylation and reactivation of silenced genes by zebularine. *J. Natl Cancer Inst.* **95**, 399–409 (2003).
- Lin, X. et al. Reversal of GSTP1 CpG island hypermethylation and reactivation of pi-class glutathione S-transferase (GSTP1) expression in human prostate cancer cells by treatment with procainamide. *Cancer Res.* **61**, 8611–8616 (2001).
- Fang, M. Z. et al. Tea polyphenol (–)-epigallocatechin-3-gallate inhibits DNA methyltransferase and reactivates methylation-silenced genes in cancer cell lines. *Cancer Res.* **63**, 7563–7570 (2003).
- Pina, I. C. et al. Psammplins from the sponge *Pseudoceratina purpurea*: inhibition of both histone deacetylase and DNA methyltransferase. *J. Org. Chem.* **68**, 3866–3873 (2003).
- Yan, L. et al. Specific inhibition of DNMT1 by antisense oligonucleotides induces re-expression of estrogen receptor- α (ER) in ER-negative human breast cancer cell lines. *Cancer Biol. Ther.* **2**, 552–556 (2003).
- Xiao, H., Hasegawa, T. & Isobe, K. Both Sp1 and Sp3 are responsible for p21waf1 promoter activity induced by histone deacetylase inhibitor in NIH3T3 cells. *J. Cell. Biochem.* **73**, 291–302 (1999).
- Marks, P. A., Miller, T. & Richon, V. M. Histone deacetylases. *Curr. Opin. Pharmacol.* **3**, 344–351 (2003).
- Jahangeer, S., Elliott, R. M. & Henneberry, R. C. β -Adrenergic receptor induction in HeLa cells: synergistic effect of 5-azacytidine and butyrate. *Biochem. Biophys. Res. Commun.* **108**, 1434–1440 (1982).
- Ginder, G. D., Whitters, M. J. & Pohlman, J. K. Activation of a chicken embryonic globin gene in adult erythroid cells by 5-azacytidine and sodium butyrate. *Proc. Natl Acad. Sci. USA* **81**, 3954–3958 (1984).
- Cameron, E. E., Bachman, K. E., Myohanen, S., Herman, J. G. & Baylin, S. B. Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. *Nature Genet.* **21**, 103–107 (1999).
- Suzuki, H. et al. A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. *Nature Genet.* **31**, 141–149 (2002).
- Yamashita, K. et al. Pharmacologic unmasking of epigenetically silenced tumor suppressor genes in esophageal squamous cell carcinoma. *Cancer Cell* **2**, 485–495 (2002).
- Belinsky, S. A. et al. Inhibition of DNA methylation and histone deacetylation prevents murine lung cancer. *Cancer Res.* **63**, 7089–7093 (2003).
- Claus, R. & Lubbert, M. Epigenetic targets in hematopoietic malignancies. *Oncogene* **22**, 6489–6496 (2003).
- Plumb, J. A., Strathdee, G., Sludden, J., Kaye, S. B. & Brown, R. Reversal of drug resistance in human tumor xenografts by 2'-deoxy-5-azacytidine-induced demethylation of the hMLH1 gene promoter. *Cancer Res.* **60**, 6039–6044 (2000).
- Karpi, A. R. & Jones, D. A. Reactivating the expression of methylation silenced genes in human cancer. *Oncogene* **21**, 5496–5503 (2002).
- Weber, J. et al. Expression of the MAGE-1 tumor antigen is up-regulated by the demethylating agent 5-aza-2'-deoxycytidine. *Cancer Res.* **54**, 1766–1771 (1994).
- Laird, P. W. The power and the promise of DNA methylation markers. *Nature Rev. Cancer* **3**, 253–266 (2003).
- Liang, G., Gonzales, F. A., Jones, P. A., Orntoft, T. F. & Thykjaer, T. Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. *Cancer Res.* **62**, 961–966 (2002).
- Mohandas, T., Sparkes, R. S. & Shapiro, L. J. Reactivation of an inactive X human chromosome: evidence for X inactivation by DNA methylation. *Science* **211**, 393–396 (1981).
- Wolf, S. F. & Migeon, B. R. Studies of X chromosome DNA methylation in normal human cells. *Nature* **295**, 667–671 (1982).
- Eversole-Cire, P. et al. Activation of an imprinted Igf2 gene in mouse somatic cell cultures. *Mol. Cell. Biol.* **13**, 4928–4938 (1993).
- Jackson-Grusby, L., Laird, P. W., Magge, S. N., Moeller, B. J. & Jaenisch, R. Mutagenicity of 5-aza-2'-deoxycytidine is mediated by the mammalian DNA methyltransferase. *Proc. Natl Acad. Sci. USA* **94**, 4681–4685 (1997).
- Carr, B. I., Rahbar, S., Asmeron, Y., Riggs, A. & Winberg, C. D. Carcinogenicity and haemoglobin synthesis induction by cytidine analogues. *Br. J. Cancer* **57**, 395–402 (1988).
- Sato, N. et al. Frequent hypomethylation of multiple genes overexpressed in pancreatic ductal adenocarcinoma. *Cancer Res.* **63**, 4158–4166 (2003).
- Yang, A. S., Estecio, M. R., Garcia-Manero, G., Kantarjian, H. M. & Issa, J. P. Comment on 'Chromosomal instability and tumors promoted by DNA hypomethylation' and 'Induction of tumors in mice by genomic hypomethylation'. *Science* **302**, 1153 (2003).

75. Lubbert, M. *et al.* Cytogenetic responses in high-risk myelodysplastic syndrome following low-dose treatment with the DNA methylation inhibitor 5-aza-2'-deoxycytidine. *Br. J. Haematol.* **114**, 349–357 (2001).
 76. Karpf, A. R., Moore, B. C., Ririe, T. O. & Jones, D. A. Activation of the p53 DNA damage response pathway after inhibition of DNA methyltransferase by 5-aza-2'-deoxycytidine. *Mol. Pharmacol.* **59**, 751–757 (2001).
 77. Jackson-Grusby, L. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* **27**, 31–39 (2001).
 78. Peterson, E. J., Bogler, O. & Taylor, S. M. p53-mediated repression of DNA methyltransferase 1 expression by specific DNA binding. *Cancer Res.* **63**, 6579–6582 (2003).
 79. Juttermann, R., Li, E. & Jaenisch, R. Toxicity of 5-aza-2'-deoxycytidine to mammalian cells is mediated primarily by covalent trapping of DNA methyltransferase rather than DNA demethylation. *Proc. Natl Acad. Sci. USA* **91**, 11797–11801 (1994).
 80. Esteller, M., Corn, P. G., Baylin, S. B. & Herman, J. G. A gene hypermethylation profile of human cancer. *Cancer Res.* **61**, 3225–3229 (2001).
 81. Cui, H. *et al.* Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science* **299**, 1753–1755 (2003).
 82. Gibbons, R. J. & Higgs, D. R. Molecular-clinical spectrum of the ATR-X syndrome. *Am. J. Med. Genet.* **97**, 204–212 (2000).
 83. Oostra, B. A. & Willemsen, R. The X chromosome and fragile X mental retardation. *Cytogenet. Genome Res.* **99**, 257–264 (2002).
 84. Ehrlich, M. The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. *Clin. Immunol.* **109**, 17–28 (2003).
 85. Nicholls, R. D., Saitoh, S. & Horsthemke, B. Imprinting in Prader–Willi and Angelman syndromes. *Trends Genet.* **14**, 194–200 (1998).
 86. Goldstone, A. P. Prader–Willi syndrome: advances in genetics, pathophysiology and treatment. *Trends Endocrinol. Metab.* **15**, 12–20 (2004).
 87. Maher, E. R. & Reik, W. Beckwith–Wiedemann syndrome: imprinting in clusters revisited. *J. Clin. Invest.* **105**, 247–252 (2000).
 88. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nature Rev. Cancer* **4**, 143–153 (2004).
 89. Soejima, H. *et al.* Silencing of imprinted *CDKN1C* gene expression is associated with loss of CpG and histone H3 lysine 9 methylation at DMR-LIT1 in esophageal cancer. *Oncogene* published online 8 March 2004 (doi:10.1038/sj.onc.1207576).
 90. Ausio, J., Levin, D. B., De Amorim, G. V., Bakker, S. & Macleod, P. M. Syndromes of disordered chromatin remodeling. *Clin. Genet.* **64**, 83–95 (2003).
 91. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
 92. Nguyen, C. T., Gonzales, F. A. & Jones, P. A. Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. *Nucleic Acids Res.* **29**, 4598–4606 (2001).
 93. Bachman, K. E. *et al.* Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. *Cancer Cell* **3**, 89–95 (2003).
 94. El-Osta, A. & Wolffe, A. P. DNA methylation and histone deacetylation in the control of gene expression: basic biochemistry to human development and disease. *Gene Expr.* **9**, 63–75 (2000).
 95. Nguyen, C. T. *et al.* Histone H3-lysine 9 methylation is associated with aberrant gene silencing in cancer cells and is rapidly reversed by 5-aza-2'-deoxycytidine. *Cancer Res.* **62**, 6456–6461 (2002).
- Acknowledgements** We thank C. Nguyen for his help with the figures, and A. Yang and J. Rice for critical reading of the manuscript. This work was supported by the National Cancer Institute and the Max Kade Foundation.
- Competing interests statement** The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Moving towards individualized medicine with pharmacogenomics

William E. Evans^{1,2} & Mary V. Relling^{1,2}

¹Department of Pharmaceutical Sciences, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

(e-mails: william.evans@stjude.org; mary.relling@stjude.org)

²University of Tennessee Colleges of Pharmacy and Medicine, and the Pharmacogenetics of Anticancer Agents Research Group, Chicago, Illinois and Memphis, Tennessee, USA

Individuals respond differently to drugs and sometimes the effects are unpredictable. Differences in DNA sequences that alter the expression or function of proteins that are targeted by drugs can contribute significantly to variation in the responses of individuals. Many of the genes examined in early studies were linked to highly penetrant, single-gene traits, but future advances hinge on the more difficult challenge of elucidating multi-gene determinants of drug response. This intersection of genomics and medicine has the potential to yield a new set of molecular diagnostic tools that can be used to individualize and optimize drug therapy.

The science of pharmacogenomics, which aims to define the genetic determinants of drug effects, has evolved over the past 50 years. The idea that genes control some drug responses was suggested in the 1950s because of the link between inheritance or ethnicity and aberrant drug responses (for a review, see ref. 1). This notion was strengthened by family and twin studies in the 1960s and 70s (ref. 2), extended by biochemical studies in the 1970s and 80s (ref. 3), and solidified by molecular genetics in the 1980s and 90s (ref. 4). Cloning and characterization of the first human gene containing DNA sequence variations (genetic polymorphisms) that influence drug metabolism did not take place until the late 1980s (ref. 5). In the decade that followed, genes responsible for many such inherited differences in drug metabolism and disposition were isolated, characterized and linked to differences in drug effects^{1,6,7}. During this period, polymorphisms in genes encoding drug transporters and targets were also identified and shown to alter drug response^{1,8}.

There remain many challenges that we have to overcome if we are to understand fully the contribution of genetic polymorphisms to inter-individual differences in drug effects and to translate this new knowledge into clinical practice. Here we discuss our view of where the field stands at present, and our vision of how it will evolve from its current state of one-drug one-gene, to one in which multiple genetic (polygenic) determinants of drug effects will be defined and used to individualize drug therapy.

A single-gene example

At present, the best-recognized and completely developed examples of genetic polymorphisms that alter drug response in humans are monogenic (single gene) traits that affect drug metabolism. One such example is that of thiopurine S-methyltransferase (TPMT) and its influence on the thiopurine drugs mercaptopurine and azathiopurine in patients who inherit non-functional *TPMT* alleles^{7,9,10}. These drugs are used clinically as immunosuppressants and to treat neoplasias. Because TPMT is the predominant inactivation pathway for these medications in haematopoietic tissues, patients who inherit a TPMT deficiency accumulate excessive concentrations of the active thioguanine nucleotides in blood cells when treated with mercaptopurine or azathiopurine. This can lead to severe and potentially life-threatening haematopoietic toxicity⁹.

TPMT-deficient patients (that is, patients with two non-functional alleles) can be treated successfully using much lower doses of thiopurines (around 5–10% of the conventional dose)^{9,10}. Clinical diagnostic tests are now available for the detection of inactivating single-nucleotide polymorphisms (SNPs) in the human *TPMT* gene, yet the routine use of TPMT genotyping to make treatment decisions is still limited^{11,12}. To some extent, this is related to the perceived high cost of genotyping, even though this process has been shown to be cost-effective¹³. In addition, resistance to change often slows the translation of science into clinical practice, a hurdle that will become even bigger when treatment decisions require information from the integration of multiple genotypes.

Where we are now

As reviewed elsewhere^{1,8} and shown in Tables 1 and 2, there is a growing list of polymorphisms found in genes encoding drug-metabolizing enzymes, drug transporters and drug targets, as well as disease-modifying genes, that have been linked to drug effects in humans. For the most part, these represent the 'low-hanging fruit' of pharmacogenetics, as they are generally monogenic and highly penetrant, with clearly recognizable drug-induced phenotypes. But this is not the case for most drug effects and treatment outcomes, which are determined by the interplay of multiple genes. This has spawned the term 'pharmacogenomics', which describes a polygenic or genome-wide approach to identifying genetic determinants of drug response, capitalizing on information from the Human Genome Project and on advances in technology (for example, high-throughput sequencing, DNA and protein microarrays, and bioinformatics). The integration of such sophisticated genomic tests with extensive phenotypic characterization of uniformly treated patients is essential if we are to define the inherited nature of most drug effects.

Genes and polymorphisms governing drug responses

Short of being able to sequence the entire genome of every patient with an aberrant drug response, how does one rationally go about identifying genetic polymorphisms that influence drug effects? The 'candidate gene' or 'candidate pathway' approach has been used to make an informed prediction of the genes in which polymorphisms might affect the disposition or response to a given drug (Fig. 1). In many cases, the known phenotypic variability in the gene product guided the search for a genetic basis for such differences.

Table 1 Examples of genetic polymorphisms that influence drug effects in humans

Protein or gene	Medications	Examples of altered drug effects
Drug-metabolizing enzymes		
CYP2C9	Non-steroidal anti-inflammatories, warfarin, tolbutamide, phenytoin	Increased anticoagulant effects of warfarin
CYP2C19	Omeprazole, mephenytoin	Peptic ulcer response to omeprazole
CYP2D6	Antidepressants, codeine, β -blockers	Increased antidepressant toxicity, decreased codeine analgesia
CYP3A4/3A5/3A7	Cyclosporin, tacrolimus, calcium channel blockers, midazolam, terfenadine, etoposide, lovastatin, tamoxifen, steroids	Decreased efficacy of tacrolimus in organ transplantation
Dihydropyrimidine dehydrogenase	Fluorouracil	Increased neurotoxicity
Glutathione transferases	Several anticancer agents	Increased response in breast cancer, more toxicity and poorer outcome in acute myeloid leukaemia
GSTM1, M3, T1		
Thiopurine methyltransferase	Mercaptopurine, thioguanine, azathioprine	Increased haematopoietic toxicity, increased risk of secondary cancers
UGT1A1	Irinotecan	Increased gastrointestinal toxicity
Drug transporters and targets		
ABCB1 (MDR-1)	Digoxin, HIV protease inhibitors, natural product anticancer drugs	Decreased CD4 response in HIV-infected patients, decreased digoxin bioavailability
β_2 -adrenergic receptor	β_2 -agonists (for example, albuterol, terbutaline)	Decreased bronchodilation
β_1 -adrenergic receptor	β_1 -antagonists	Decreased cardiovascular response to β_1 -antagonists
Gs protein α	β -blockers (for example, metoprolol)	Decreased antihypertensive effect
ALOX5	Leukotriene receptor antagonists	Lower changes in FEV ₁ (forced expiratory volume)
Serotonin transporter (5-HTT)	Antidepressants (for example, fluoxetine)	Decreased clozapine effects, antidepressant response

This table is intended to be representative and not comprehensive. Primary literature citations are available in recent reviews^{1,6,7,8,30}.

For example, inter-individual differences in response to the anti-hypertensive drug debrisoquine and in the metabolism of the oxytotoxic drug sparteine led to family studies that revealed that the pharmacodynamic effects of these drugs are inherited^{14,15}. This led to surveys of human livers to identify the enzyme responsible for debrisoquine metabolism and to isolation of the gene responsible — *CYP2D6*. Common polymorphisms that cause phenotypic variability in *CYP2D6* activity were then identified^{5,16}. Although single-gene defects (for example, *CYP2D6* and *TPMT*) can have a strong effect on their drug substrates, most of the phenotypic variability in drug response remains unexplained, despite numerous efforts to interrogate candidate genes and pathways.

The candidate-gene approach can fail for several reasons, one of which could be that other mechanisms alter protein function (for example, post-translational modifications). Even if the selected gene is important, it may be technically difficult to identify the functionally important polymorphisms, which can include promoter or enhancer polymorphisms, gene duplications, synonymous coding SNPs that affect transcript stability, or intronic SNPs that cause splice variants that create early stop codons. Of course, SNPs can also cause significant amino-acid substitutions^{1,4,6–8}.

'Pathways' of genes may be more important than individual genes, with the effects of polymorphisms in networks of genes acting together to create a single phenotype. For example, high oxidative capacity in a cytochrome P450 coupled with low conjugating activity

in an inactivating glucuronosyltransferase could result in greater exposure to toxic oxidative metabolites. Alternatively, polymorphisms in genes on common pathways may offset each other, so that the effect of one gene is not easily detected. For example, more black people than white have gene polymorphisms associated with high expression of *CYP3A5*, a member of the cytochrome P450 family and an important enzyme for metabolizing many drugs. But the same drugs are often excreted by p-glycoprotein, and more white people than black have gene polymorphisms associated with high levels of p-glycoprotein⁸. Thus, polymorphisms in *CYP3A5* and p-glycoprotein may counterbalance each other and reduce overall population variability in response to drugs that are substrates for both.

Of course, the candidate-gene approach may also fail because the wrong genes have been targeted. Genome-wide approaches, such as gene-expression arrays, genome-wide scans or proteomic assays, can assist in the identification of as-yet-unrecognized candidate genes (Fig. 1). Perhaps this can be accomplished by detecting genes whose expression differentiates drug responders from non-responders (or those that are toxic from those that are not)¹⁷, genomic regions with a paucity of heterozygosity in responders compared with non-responders¹⁸, or proteins whose abundance differentiates drug responders from non-responders¹⁹. Expression-array and proteomic approaches have the advantage that the level of the signal may directly reflect functional variation. But both have the disadvantages that experiments are highly influenced by choice of

Table 2 Examples of genetic polymorphisms in disease- or treatment-modifying genes that alter drug response in humans

Protein or gene	Disease or response association	Medications	Impact of polymorphism on drug effect/toxicity
Adducin	Hypertension	Diuretics	Decreased myocardial infarction or strokes
APOE	Atherosclerosis progression, ischaemic cardiovascular events	Statins (for example, simvastatin)	Enhanced survival prolongation
HLA	Toxicity	Abacavir	Increased risk of hypersensitivity reaction
CETP	Atherosclerosis progression, HDL-cholesterol levels	Statins (for example, pravastatin)	Slowing atherosclerosis progression
HERG, KvLQT1, Mink, MIRP1	Congenital long QT syndrome	Erythromycin, terfenadine, cisapride, clarithromycin, quinidine	Increased risk of drug-induced Torsade de Pointes
MGMT	Glioma	Carmustine	Increased response of glioma to carmustine
Parkin	Parkinson's disease	Levodopa	Clinical improvement and L-dopa-induced dyskinesias
Prothrombin and factor V	Deep vein thrombosis and cerebral vein thrombosis	Oral contraceptives	Increased deep vein and cerebral vein thrombosis risk with oral contraceptives
Stromelysin-1	Atherosclerosis progression	Statins (for example, pravastatin)	Reduction in cardiovascular events by pravastatin

This table is intended to be representative and not comprehensive. Primary literature citations are available in recent reviews^{1,6,7,8,30}.

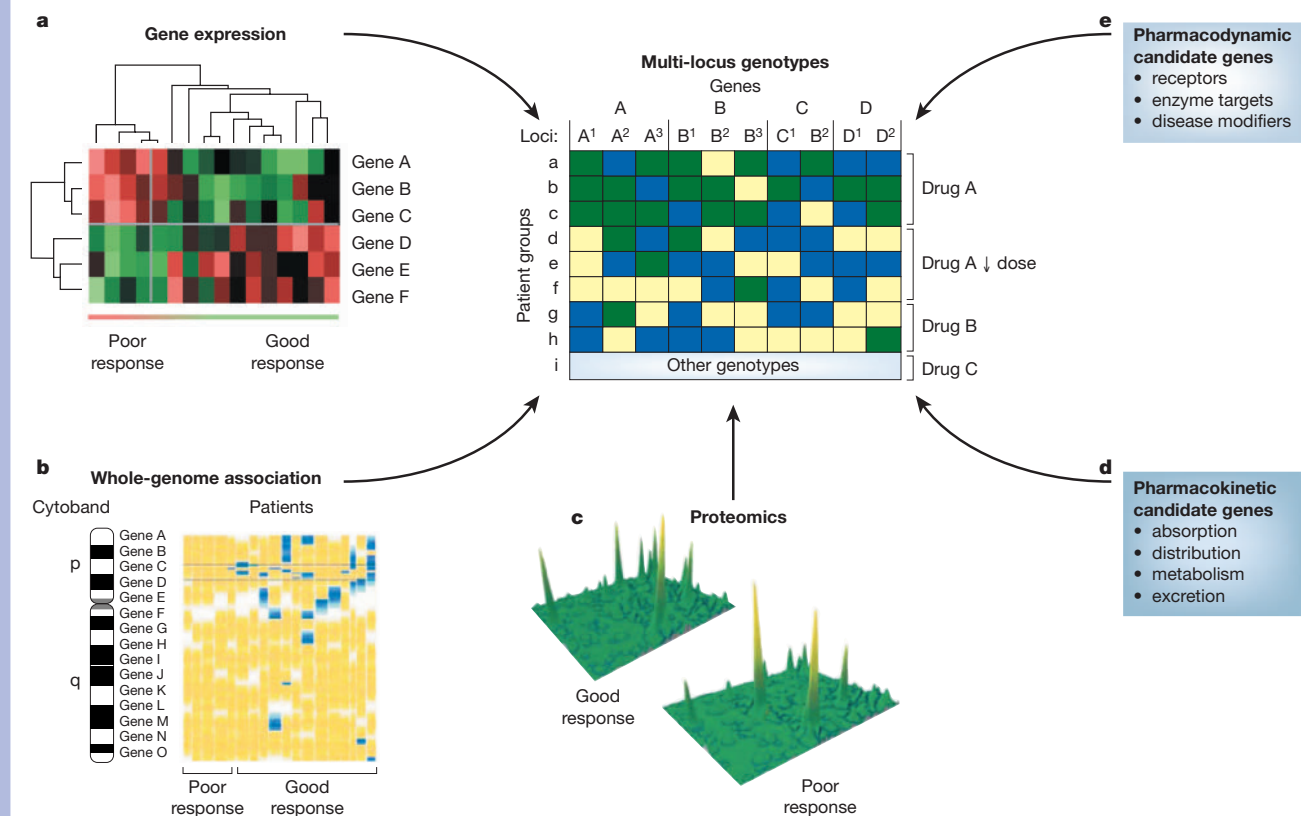


Figure 1 Identifying genes influencing polygenic drug responses. **a–c**, Three genome-wide approaches — gene-expression analysis, genome-wide scans and proteomics — can be used to identify potential candidate genes that influence a specific drug response (for example, the promotion of anti-leukaemic effects). **a**, DNA microarray analysis of cancer cells for 6 genes (rows) in 16 patients (columns), 4 with a poor response and 12 with a good response. **b**, SNP haplotype map showing 16 gene loci on one chromosome for 5 patients with a poor response and 13 with a good response. **c**, LC–MS (liquid chromatography/mass spectrometry) analysis of plasma or tumour tissue to identify differences in proteins (see yellow peaks) between good and poor responders. **d, e**, The conventional ‘candidate gene’ approach, based on clinical pharmacology studies of proteins (for example, receptors) and pathways known to be involved in a drug’s pharmacokinetic or pharmacodynamic response. The middle panel represents a subset of the final product, a panel of multiple genes and loci that collectively segregate patients into groups who respond best to one of several drugs or drug doses.

tissue, which may or may not reflect the tissue of concern for toxicity and response, that false negatives are possible (with changes in function not reflected by levels of messenger RNA or protein) and, of course, that the large number of transcripts or proteins that can be interrogated leads to the possibility of false-positive discoveries.

Once a gene or its product has been implicated in a drug response, large-scale molecular epidemiological association studies (*in vivo* or *in vitro* with human tissues), biochemical functional studies and pre-clinical animal models of candidate gene polymorphisms can be used to further establish genetic variability as a determinant of specific drug effects (Fig. 2).

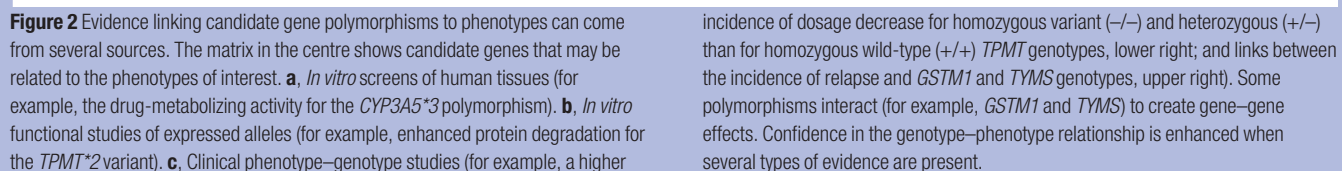
Enhancing drug discovery and development

There was an initial exuberance that the Human Genome Project would reveal many novel targets against which new medications could be developed. Although this promise has yet to be realized, there remains great potential for the human genome to yield new insights into the pathogenesis of human diseases and to reveal new strategies for their prevention or treatment. Pharmacogenomics may enhance drug discovery and development in two ways: by the identification of drug targets and by subpopulation-specific drug development.

For example, genomics can be used to identify new targets through the discovery of genes that are under- or overexpressed in cancer cells that are sensitive to anticancer agents compared with those that are resistant. The products of such overexpressed genes represent plausible targets for inhibitors that could reverse the drug-

resistance phenotype. Another application is the use of gene expression in target tissues (for example, cancer cells) to ascertain the effects of chemotherapy and how cells respond to treatment with single drugs or drug combinations^{17,20}. A recent extension of this strategy is to use gene-expression signatures that connote the desired drug effects²¹ (for example, cancer cell differentiation) as the read-out for high-throughput screening of candidate compounds²².

Another use of pharmacogenomics is to identify genetic polymorphisms that predispose patients to adverse drug effects that, although they may occur in only a small subset of the people treated with a new medication, are sufficiently toxic to jeopardize further development of the drug for all patients. One approach is to obtain genomic DNA from patients entered on large Phase III clinical trials of a new agent, and then retrospectively to search for genetic polymorphisms that predispose a small subset to severe toxicities, should they arise. The hope, at the outset, is that no such toxicities would emerge for new agents under development, but should severe toxicity occur in a very small percentage of patients who could be prospectively identified based on genotype, then an otherwise efficacious new drug might be ‘saved’ from abandonment during development, or from withdrawal after approval and widespread use. For example, patients at high risk of abacavir hypersensitivity can be identified on the basis of their HLA-B genotypes²³, but it remains unclear whether prospective screening of HLA genotypes is cost-effective with current technology. More life-threatening toxicities would probably justify such testing with new medications that have unique pharmacological properties.



Given the potential value of knowing all the possible factors that influence the effects of new agents, it is likely that pharmacogenomics will have an increasingly important role in drug discovery and development. Pharmacogenomics has been used to 'kill' new drugs whose metabolism depends primarily on an enzyme that exhibits genetic polymorphism. If such screening had been available in the 1950s, mercaptopurine and azathioprine may never have been developed, and the hundreds of thousands of patients who have benefited from these medications would have missed out. Science now offers better alternatives than simply discarding a potentially effective drug because it interacts with a protein that exhibits genetic polymorphism.

Translating pharmacogenomics to the clinic

First, we have spent decades educating clinicians to start treatment with the default 'average dose'. Individualizing dosages, even based on easily assessed patient characteristics (such as age or renal function), has not been widely embraced by the medical or pharmaceutical communities. There is resistance to relying on tests for every medical decision, and a 'trial and error' approach to drug dosing has become widely accepted. Not only does pharmacogenomics require a laboratory test, it also requires an interpretation of genotypes, which will probably require clinicians to receive further training in molecular biology or genetics.

Also, it is difficult to conduct definitive clinical pharmacogenomic studies to prove that individualization of drug therapy on the basis of genetics improves clinical outcomes. Given the multigenic nature of most drug effects, the difficulty in controlling for non-genetic confounders (for example, drug interactions, diet and smoking) and the lack of funding for large-scale pharmacogenomic studies with adequate follow-up, it is not surprising that there is a

lack of evidence available to catalyse a change in clinical practice. Our enthusiasm for advancing molecular technology and defining the human genome has not yet been matched by a willingness to incorporate this technology and knowledge into well-controlled and monitored clinical trials.

In fact there are remarkably few examples of genetic technology that have been validated to the point where they can be used in clinically licensed and regulated laboratories. The diverse range of mechanisms that account for genetic polymorphisms (SNPs, insertions/deletions, splice variants and so on) means that providing definitive results for even a single gene is a tremendous challenge, as has been demonstrated for *BRCA1* and *BRCA2*, which are implicated in breast cancer, and *CFTR*, which has a role in cystic fibrosis²⁸. Ultimately, testing for pharmacogenetic polymorphisms must overcome the same technical hurdles that govern other molecular diagnostics. The good news is that a given genotype needs to be determined only once, unlike a measure of renal function.

Future directions

It seems likely that advances in technology will drive down the cost of genotyping sooner than science and medicine will be able to establish definitive polygenic models for optimizing drug therapy. But as pharmacogenomic strategies become a routine part of drug discovery and development, such models should emerge from large-scale clinical trials and follow-up of new medications. Public initiatives — such as the Pharmacogenetics Research Network (<http://www.PharmGKB.org>), funded by the National Institutes of Health — coupled with other academic and industry initiatives, will remain a driving force behind this science²⁹. Given the millions of human genomic polymorphisms, the best way forward is probably through the replication of large-scale clinical trials of uniformly treated and evaluated patients. These trials should incorporate comprehensive and rigorous pharmacogenomic studies, coupled with preclinical experimental models that reinforce genotype–phenotype clinical associations. Incorporation of genetic tests into clinical trials may not be deemed ‘innovative’ by conventional peer review, but will be necessary for clinical progress.

If the use of medication is to evolve from a ‘trial and error’ approach to individualized therapy using genetics, at least two aspects of health care need to change. We must introduce protection against the misuse of genetic information and accept the added costs that may be incurred during the transition to genetically guided decisions about drug therapy. In the long run, decreasing the frequency of adverse drug effects and increasing the probability of successful therapy will probably lower the cost of health care. Pharmacogenomics has the potential to facilitate this process by translating knowledge of human genome variability into better therapeutics. □

doi:10.1038/nature02626

1. Evans, W. E. & Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
2. Vesell, E. S. Pharmacogenetic perspectives gained from twin and family studies. *Pharmacol. Ther.* **41**, 535–552 (1989).
3. Guengerich, F. P. *et al.* Twenty years of biochemistry of human P450s: purification, expression, mechanism, and relevance to drugs. *Drug Metab. Dispos.* **26**, 1175–1178 (1998).
4. Meyer, U. A. & Zanger, U. M. Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annu. Rev. Pharmacol. Toxicol.* **37**, 269–296 (1997).
5. Gonzalez, F. J. *et al.* Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* **331**, 442–446 (1988).
6. Evans, W. E. & Johnson, J. A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genet.* **2**, 9–39 (2001).
7. Weinsztilboum, R. Inheritance and drug response. *N. Engl. J. Med.* **348**, 529–537 (2003).
8. Evans, W. E. & McLeod, H. L. Pharmacogenomics — drug disposition, drug targets, and side effects. *N. Engl. J. Med.* **348**, 538–549 (2003).
9. Evans, W. E. *et al.* Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine. *J. Clin. Oncol.* **19**, 2293–2301 (2001).
10. Evans, W. E., Horner, M., Chu, Y. Q., Kalwinsky, D. & Roberts, W. M. Altered mercaptopurine metabolism, toxic effects, and dosage requirement in a thiopurine methyltransferase-deficient child with acute lymphocytic leukemia. *J. Pediatr.* **119**, 985–989 (1991).
11. Marshall, E. Preventing toxicity with a gene test. *Science* **302**, 588–590 (2003).
12. Abbott, A. With your genes? Take one of these, three times a day. *Nature* **425**, 760–762 (2003).
13. Phillips, K. A., Veenstra, D. L., Oren, E., Lee, J. K. & Sadee, W. Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *J. Am. Med. Assoc.* **286**, 2270–2279 (2001).
14. Eichelbaum, M., Spannbrucker, N. & Dengler, H. J. N-oxidation of sparteine in man and its interindividual differences. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **287**, R94 (1975).
15. Mahgoub, A., Idle, J. R., Dring, L. G., Lancaster, R. & Smith, R. L. Polymorphic hydroxylation of Debrisoquine in man. *Lancet* **2**, 584–586 (1977).
16. Ingelman-Sundberg, M., Oscarson, M. & McLellan, R. A. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol. Sci.* **20**, 342–349 (1999).
17. Cheok, M. H. *et al.* Treatment-specific changes in gene expression discriminate *in vivo* drug response in human leukemia cells. *Nature Genet.* **34**, 85–90 (2003).
18. Buetow, K. H. *et al.* High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl Acad. Sci. USA* **98**, 581–584 (2001).
19. Liotta, L. A., Kohn, E. C. & Petricoin, E. F. Clinical proteomics: personalized molecular medicine. *J. Am. Med. Assoc.* **286**, 2211–2214 (2001).
20. Golub, T. R. Mining the genome for combination therapies. *Nature Med.* **9**, 510–511 (2003).
21. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
22. Evans, W. E. & Guy, R. K. Gene expression as a drug discovery tool. *Nature Genet.* **36**, 214–215 (2004).
23. Mallal, S. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
24. Kuehl, P. *et al.* Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nature Genet.* **27**, 383–391 (2001).
25. Holden, C. Race and medicine. *Science* **302**, 594–596 (2003).
26. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
27. Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
28. Eccles, D. M. Genetic testing for BRCA1 mutation in the UK. *Lancet* **361**, 178–179 (2003).
29. Altman, R. B. *et al.* Indexing pharmacogenetic knowledge on the World Wide Web. *Pharmacogenetics* **13**, 3–5 (2003).
30. Goldstein, D. B., Tate, S. K. & Sisodiya, S. M. Pharmacogenetics goes genomic. *Nature Rev. Genet.* **4**, 937–947 (2003).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Oncogenomics and the development of new cancer therapies

Robert L. Strausberg¹, Andrew J. G. Simpson², Lloyd J. Old² & Gregory J. Riggins³

¹Department of Mammalian Genomics, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 2085, USA (e-mail: RLS@tigr.org)

²Ludwig Institute for Cancer Research, 605 Third Avenue, New York, New York 10158, USA

³Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, Maryland 21224, USA

Scientists have sequenced the human genome and identified most of its genes. Now it is time to use these genomic data, and the high-throughput technology developed to generate them, to tackle major health problems such as cancer. To accelerate our understanding of this disease and to produce targeted therapies, further basic mutational and functional genomic information is required. A systematic and coordinated approach, with the results freely available, should speed up progress. This will best be accomplished through an international academic and pharmaceutical oncogenomics initiative.

Cancer can result from a multitude of genetic alterations. These range from point mutations to insertions and deletions to chromosomal translocations in the tumour cell's DNA^{1–3}. The completion of the human genome sequence^{4,5}, rapid improvements in our understanding of the molecular basis of cancer, and the introduction of new technologies for assessing genomic, transcriptional and proteomic changes provide new opportunities for addressing the practical challenges of cancer. In particular, prospects for the development of new therapeutics appear to be excellent.

Important cancer-causing mutations, genomic rearrangements and transcriptional changes could be fully catalogued during the next decade, judging by the success of recent oncogenomics initiatives. Coordinated international research efforts analogous to the Human Genome Project (HGP) have the potential to produce an integrated database of cancer-causing mutations, and of the subsequent transcriptional and translational alterations. Such a database could directly identify targets for therapeutic agents.

The clinical responses seen with the first generation of targeted therapeutic agents, such as the kinase inhibitor Gleevec⁶ and the monoclonal antibodies Rituxan⁷, Herceptin⁸ and Avastin⁹, have raised expectations that in future there will be an arsenal of targeted cancer therapies of high efficacy and specificity at our disposal^{10,11}. Before us is an extraordinary opportunity to contribute to the development of new therapies, but how best to direct and coordinate our efforts is a question we need to investigate further and is the subject of this review.

The importance of genomic technology

The use of automated sequencing technology and the completion of the human genome have accelerated the identification of mutations and alterations in cancer genomes that might be 'targetable' by various therapies. The most easily 'banked' data — and the most definitive of molecular biology data forms — are DNA sequence information. But significant challenges remain: finding small alterations (often just a difference of a single base pair) in the genome is extremely difficult, as is making such studies comprehensive across the many types of cancer and ensuring that large amounts of data are reported in a meaningful fashion. Nevertheless, we are optimistic that a comprehensive mutational database is possible for most human cancers.

Many parts of the genome are now being resequenced to identify somatic alterations in exons, and several important cancer mutations have been identified using this strategy^{12–14}. Furthermore, advances in DNA sequencing technologies may further reduce the cost and increase the efficiency of obtaining cancer genome sequences.

The availability of the human genome sequence has made alterations such as chromosomal aberrations, translocations, deletions, amplifications and methylation easier to locate^{2,15–20}. The new sequence-based approaches, such as digital karyotyping^{21,22}, a method for detecting DNA copy number on a genomic scale, allow sequence amplifications and deletions to be discovered in a high-throughput manner. And the recently introduced bacterial artificial chromosome (BAC)-end sequencing²³ identifies chromosomal breakpoints in a comprehensive, high-resolution manner. The challenge is to gain the best from each technology so that an efficient and comprehensive analysis of each cancer can be achieved.

Although much effort is focused on individual cancer types, the sequence-based nature of many approaches, with the genome as a point of reference, allows ready database compilation. Databases have already been established for mutations in individual genes such as p53 (ref. 24). Such initiatives could eventually serve as a focal point or model for coordinated community-wide compilations of cancer-related mutations^{12,13,25} and the prospect of a more comprehensive approach.

It is not just genomic technology that has improved — methods for detecting messenger RNA have too. Microarrays make major contributions to this because they provide a cost-effective means of assessing and comparing mRNA levels in multiple samples^{26–29}. Indeed, studies using this technology have suggested that transcription profiles allow the molecular classification of cancers, as well as insight into the biology of tumour progression^{30–32}. This raises the possibility of novel approaches for diagnosing cancers and for predicting response to therapy. In addition, sequence-based approaches for gene tagging — for example, serial analysis of gene expression (SAGE)^{33–35}, massively parallel signature sequencing³⁶ and expressed sequence tags (including ORESTES)³⁷ — provide data sets that facilitate and complement the microarray approaches.

The large-scale transcript sequencing projects, such as those of the United States and Brazil, have resulted in the establishment of and presentation of data on the Cancer

Genome Anatomy Project website (<http://cgap.nci.nih.gov>), a key reference for the definition of human gene expression in normal tissues and tumours³⁸. Although sequence-based approaches such as SAGE are currently less well suited to the analysis of large numbers of samples, sequencing facilitates the discovery of new genes, thereby serving as a platform for the design of microarrays. In addition, the precise digital nature of DNA sequencing is particularly well suited to transcript quantification and the development of community-wide databases^{34,38}.

The continued improvements in DNA sequencing have increased the cost-effectiveness of transcript tagging, and have made it easier to obtain more information about rarer transcripts. In addition, the use of longer SAGE tags has improved the linkage of tags to specific genes and within the genome³⁹. Although still expensive, full-length complementary DNA sequencing remains the 'gold standard' for both gene identification and the definition of gene structure^{40,41}. Full-length cDNA sequences will be crucial for the eventual comprehensive documentation of alternative splice forms^{42–44} in cancer, and for the discovery of their biological roles in the initiation and progression of the disease. Although efforts have begun⁴², these transcripts have not been comprehensively catalogued in normal cells and cancer cells, and a complete picture has yet to emerge. Increasing emphasis on full-length cDNA sequencing promises to greatly increase our knowledge of splice variants in cancer and offers the potential to identify new cancer-specific targets.

An important goal for oncogenomics is the integration of complete genome analyses and transcriptome analyses so that accurate models of the molecular basis of cancer can be built. This step will depend first and foremost on the further decrease in cost and increase in scale of the analytical technologies. But it will also depend on how well these activities are organized: the new databases and analytical tools must be aimed at providing integrated views of the changes in cancer within cells and across diverse cancers.

The need for enhanced informatics tools

Database development and analysis tools were vital for the HGP. They will be equally important for the development and population

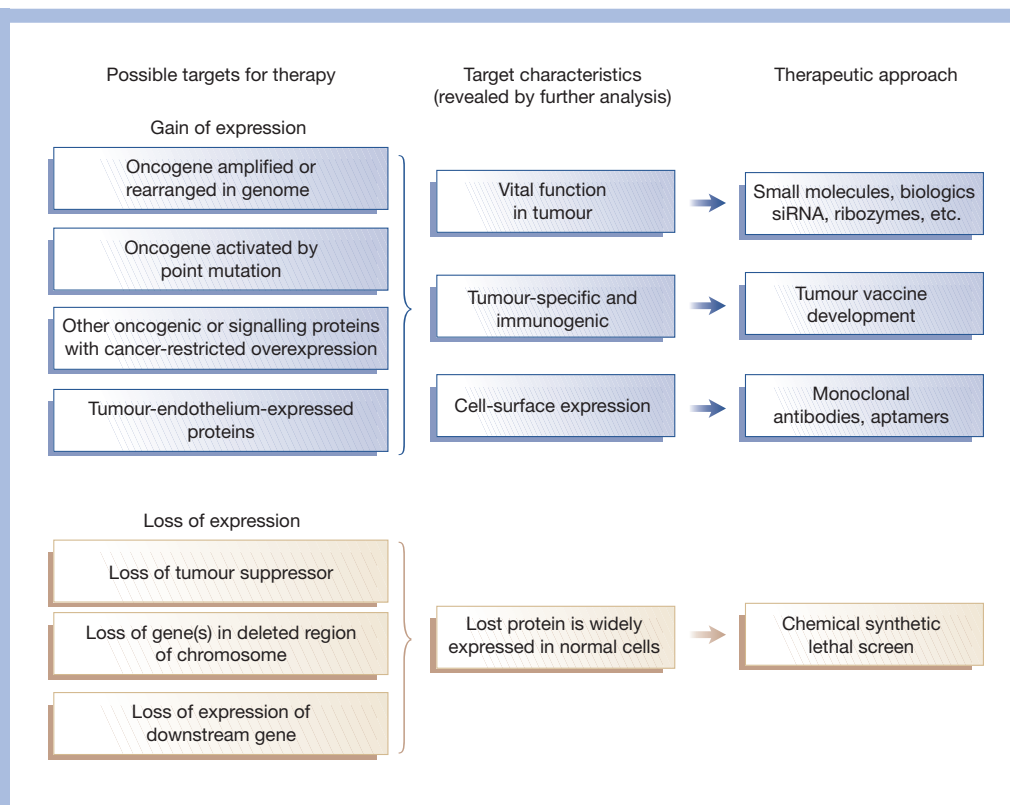
of a database of the molecular biology of cancer. The goal is to incorporate various types of genomic data (such as mutations, transcriptional changes and cytogenetic events) in a common database with standard terms (for example, using a common scheme for tumour classification). Already, cancer genome projects have been established that provide a starting point for this^{12,34,38,45}. Given the existing template provided by the human genome sequence, it is feasible to display diverse data sets (especially those with sequence-based information) in a common framework that is accessible to all. To provide a complete view of cancer, it is important that the database contains not only information about molecular changes in cancers, but also data documenting the absence of changes in cancer, such as resequenced genes in cell lines or tumours that are found to be wild type. This will provide a more complete view of tumours and will help to avoid duplication of resequencing efforts. Similar to the HGP, the cancer genome project would need a set of standards for the data.

Discovering cancer targets

Central to expectations of accelerated target discovery is the perception that genome, transcriptome and proteome analyses will lead to the identification of molecules against which cancer therapeutics might be targeted (Fig. 1). Among many possible approaches, the three that are currently receiving the most attention are: (1) small-molecule inhibitors of oncogenic signals, (2) antibodies to surface components and intercellular communicating factors, and (3) molecularly defined vaccines.

For the first of these options, obvious targets are oncogenes, such as kinases^{46,47}, because of their causal involvement in tumorigenesis. The development of targeted chemical and biological agents that reduce their oncogenic function has been successful^{48,49}. Oncogenes can be putatively identified by the amplification of the gene that encodes them, the over-representation of their transcript or protein in tumours, chromosomal translocation, and/or the cancer-related occurrence of activating missense mutations in their gene sequences². The best example of a targeted, small-molecule oncogene inhibitor that interrupts a crucial signalling pathway is imatinib (better known by its commercial name, Gleevec), which is already in clinical use.

Figure 1 Types of target identified through comprehensive genome and transcriptome analysis of cancer cells. Most genetic targets will result from a 'gain of expression', as in the case of genes that are activated by either a point mutation or a chromosomal rearrangement that creates a new protein sequence. Targets can also be proteins with a normal sequence that are pathologically overexpressed by genomic amplifications, promoter mutations or upstream pathway activation. An example is the discovery of epidermal growth factor receptor (*EGFR*) genomic amplification, leading to its overexpression and increased kinase activity, which can be targeted by small molecules. Designing compounds that are toxic to cells on the basis of 'loss of expression' might be an alternative way to target cancer cells; this approach could be adopted in the case of missing tumour-suppressor proteins, a genetic event that is common in tumour development.



It was the observation of a ubiquitous chromosomal translocation in chronic myelogenous leukaemia that first prompted the development of Gleevec. Although Gleevec was initially developed as an inhibitor of the bcr-abl kinase⁶, it is now known to inhibit a select family of tyrosine kinases, including KIT and PDGFR, in a variety of tumour types⁵⁰. The application of Gleevec to a specific, oncogene-dependent form of sarcoma, known as gastrointestinal stromal tumour (GIST), had spectacular anti-tumour effects and provided proof that both haematopoietic and solid tumours are amenable to targeted therapy^{51,52}.

The frequency of relapse because of resistance currently limits the effectiveness of this treatment. The need for other treatments is stimulating the search for similarly effective molecules that might be simultaneously applied, as in the anti-HIV cocktail^{53,54}.

Part of the success of Gleevec can be attributed to the intensity and productivity of the academia-industry interaction that took place throughout its development as an anti-cancer agent. This new and effective cancer treatment would not be available today without this cooperative interaction^{6,52}, nor would its wider application to include the treatment of the solid tumour GIST have been achieved^{55,56}. Furthermore, the first treatment for Gleevec-resistant patients is now emerging from a similar academia-industry interaction with the development of the multi-kinase inhibitor SU11248 (ref. 57).

Transcriptome and genome data also identify targets that are potentially suitable for immunotherapeutic intervention. Among the genes that are most actively sought are those that have their expression restricted to one or more non-essential tissues or cells, such as the breast, the prostate or melanocytes, irrespective of disease state⁵⁸⁻⁶¹. These genes are potential targets for vaccines or antibodies, because the targeted destruction of the remnants of an organ, such as the prostate, following its surgical removal is clinically acceptable irrespective of the disease state of the individual cells. As potential targets, these molecules have the great advantage of typically being present at high levels in tumours. In addition, new genes are being discovered that are expressed in a range of tumours but are restricted in their normal expression to one or more non-essential or embryonic tissues distinct from that in which the tumour is found. A particularly striking example is the growing number of cancer genes in the testis that are expressed only in tumours and in human gametes⁶²⁻⁶⁴. Although the contribution of these genes to the process of tumorigenesis is not yet known, they are important targets of potential cancer vaccines⁶⁵.

There are not, as yet, any molecular, therapeutic cancer vaccines for clinical use, although clinical trials are underway^{66,67}. However, antibodies are already making a substantial contribution to therapy⁶⁸⁻⁷⁰. The monoclonal antibody Herceptin, for example, is targeted to the epidermal-growth-factor-like receptor encoded by the *HER-2/neu* gene, which is overexpressed in 20–30% of breast cancer patients. Rituxan is targeted to CD20 (which is selectively expressed on B cells) and is used for the treatment of patients with non-Hodgkin's lymphoma both as single-agent therapy and in combination with chemotherapy⁷¹. Providing further optimism for the application of antibodies to cancer intervention is the recent introduction of Avastin⁹, the first targeted anti-angiogenesis agent, now approved as a treatment for metastatic colon cancer⁷².

Herceptin, Rituxan and Avastin all function as naked antibodies. But antibodies can also be used to deliver toxic compounds to the tumour. Examples of this approach have already reached the clinic. Zevalin, a derivative of Rituxan ligated to the radioisotope ⁹⁰Y, is the first radio-labelled monoclonal antibody to be approved for the treatment of cancer⁷³. Another example is Mylotarg, a monoclonal antibody used for the treatment of acute myeloid leukaemia. It targets CD33 and is coupled to the toxin calicheamicin, which initiates apoptosis⁷⁴.

From genomic discovery to clinical utility

Mutations in certain tumour suppressors, such as p53 and APC, occur in many cancer types at high frequency. This information has raised hopes that targeted therapies could work on a broad spectrum

of tumours. Although exploitation of these targets has proved to be quite challenging, several creative approaches are being explored, such as the use of selective replication of viruses in cancer cells, RNA interference or ribozymes⁷⁵⁻⁷⁸. For example, viruses have been engineered that selectively replicate in cells that have a loss of p53 function⁷⁵.

In the meantime, the most commonly used treatment involves the targeting of oncogenes, in particular kinases. But alterations in specific kinase genes are seen only in a subset of tumours. For example, colon cancer has a broad spectrum of low-frequency mutations in kinases, each of which could be a potential target^{13,14}. This molecular heterogeneity complicates the development of universally applicable therapies. The development of vaccines and antibody therapy against an array of cancers is similarly hampered by potential targets being found only in subsets of any tumour type. An individual target is rarely present in as many as half of the tumours that arise in a given tissue⁶².

Although a piecemeal approach to therapy may be unavoidable at present, a comprehensive genomic approach to the discovery of molecular alterations may identify unifying features that could result in broader-range therapies. Such common features could provide a means of targeting rare tumours or tumours more common in distinct geographical areas in a cost-effective manner. For example, in the case of kinase mutations in colon cancer, although no single mutation has been found in all of these cancers, a high proportion of cancers do have a kinase mutation, sometimes at the corresponding amino-acid position in different genes — for example, in the autoinhibitory activation loop¹³.

It is highly likely that targetable mutations (or other molecular changes) common to many cancers will be found and will open the way for the targeting of different types of tumour (including rare tumours that otherwise might not receive sufficient attention) that can be inhibited by the same agent, as in the case of Gleevec. A key to progress in this area is the establishment of assays that can indicate potential success in the clinic.

Tools and assays for targeting molecular changes in cancer

To discover useful small-molecule inhibitors once mutations or other alterations have been identified in cancers, several approaches (Fig. 2) are available. One clever strategy is to use two cell lines that are identical (isogenic) except for a single cancer-causing mutation for screening. For example, Kinzler and colleagues, who developed this approach, compared cells with and without a *K-ras* mutation⁷⁹. The two cell lines were labelled with different fluorescent markers, then mixed together and allowed to grow in the presence of small-molecule libraries. The relative growth rates of the cells were monitored using fluorescence spectroscopy. If any of the small molecules added inhibited the growth of the cancer-causing cell line, but not the other line, then its properties were explored further.

To be able to screen for inhibitors, scientists in research laboratories of even moderate size need to have access to materials such as diverse small-molecule libraries both for drug development and for the analysis of gene function in model systems or cell lines (Fig. 3).

Although at an early stage, ways of accessing chemical libraries are being introduced, and there are strong indications that this process will continue. One such programme is the National Cancer Institute (NCI) Initiative for Chemical Genetics (ICG), which includes support for ChemBank (<http://chembank.med.harvard.edu>)⁸⁰. This programme aims to systematically identify perturbational small molecules for each cancer-related protein coded in the human genome. This should help to better define regulatory pathway networks in cancer, thereby providing valuable information for the development of cancer treatments⁸¹. Coupled with the National Institutes of Health (NIH) newly announced goal — part of its Roadmap initiative (<http://nihroadmap.nih.gov>) — to apply small-molecule drug discovery on a genome scale⁸², as well as the increasing role of the National Human Genome Research Institute in translational

research⁸³, the prospects of greater access to molecules for screening look good. Other established efforts, such as the NCI's Rapid Access to Investigational Drugs (RAID) programme (http://dtp.nci.nih.gov/docs/raid/raid_index.html), are providing opportunities for academic (and industrial) researchers to have an active role in clinical development. Moreover, the NIH Roadmap initiative is now strongly advocating major academic participation in many areas of translational research.

Although much of the emphasis has been placed by researchers on the screening of synthetic molecules created by combinatorial chemistry, it should be remembered that the majority of clinically successful small molecules for the treatment of cancer are based on natural compounds⁸⁴. Most of the world's biodiversity exists in developing countries, which highlights the need to work internationally to improve cancer therapy⁸⁵.

The availability of small-molecule libraries is only a starting point. Functional data are needed for each target, testing in relevant animal models is required, and potential treatments (leads) have to be improved, re-tested and refined by medicinal chemists. Thereafter, the problems of good manufacturing practice (GMP) production and regulatory approval have to be addressed before beginning to negotiate issues such as toxicology, pharmacokinetics and access to tumour tissue.

Similar difficulties need to be overcome for therapeutic antibodies and vaccines to be developed. For either form of immunotherapy, the distribution of the potential target protein throughout the body will need to be determined, and for this, monoclonal antibodies will be used as investigative reagents. Although such antibodies can be raised using DNA or peptides, the use of whole protein has historically proved to be most reliable. But this requires the routine production of at least small quantities of such proteins, a process that is currently low throughput, although projects aimed at the large-scale generation (within the academic domain) of comprehensive sets of human proteins have been planned (FLEXGene)⁸⁶.

In addition, the analysis of naturally occurring immune responses to the potential cancer-causing targets is important in identifying those that are most likely to be immunogenic and protective. Even after the suitability of a target has been confirmed, further challenges lie ahead: the availability of experimental therapeutic reagents, in the form of antibodies or recombinant proteins, is even more limited than it is for small-molecule inhibitors. At present, antibodies are produced in mice and then adapted for human use ('humanized'). Alternative strategies include the use of phage display libraries for antibody generation^{87,88}, as well as selectively binding peptides^{89,90}. The production of the quantity (grams) required for initial human trials is a major undertaking. And the production of gram quantities of highly purified recombinant proteins for vaccines, under GMP conditions, can at present only be pursued on a one-by-one basis in purpose-built facilities.

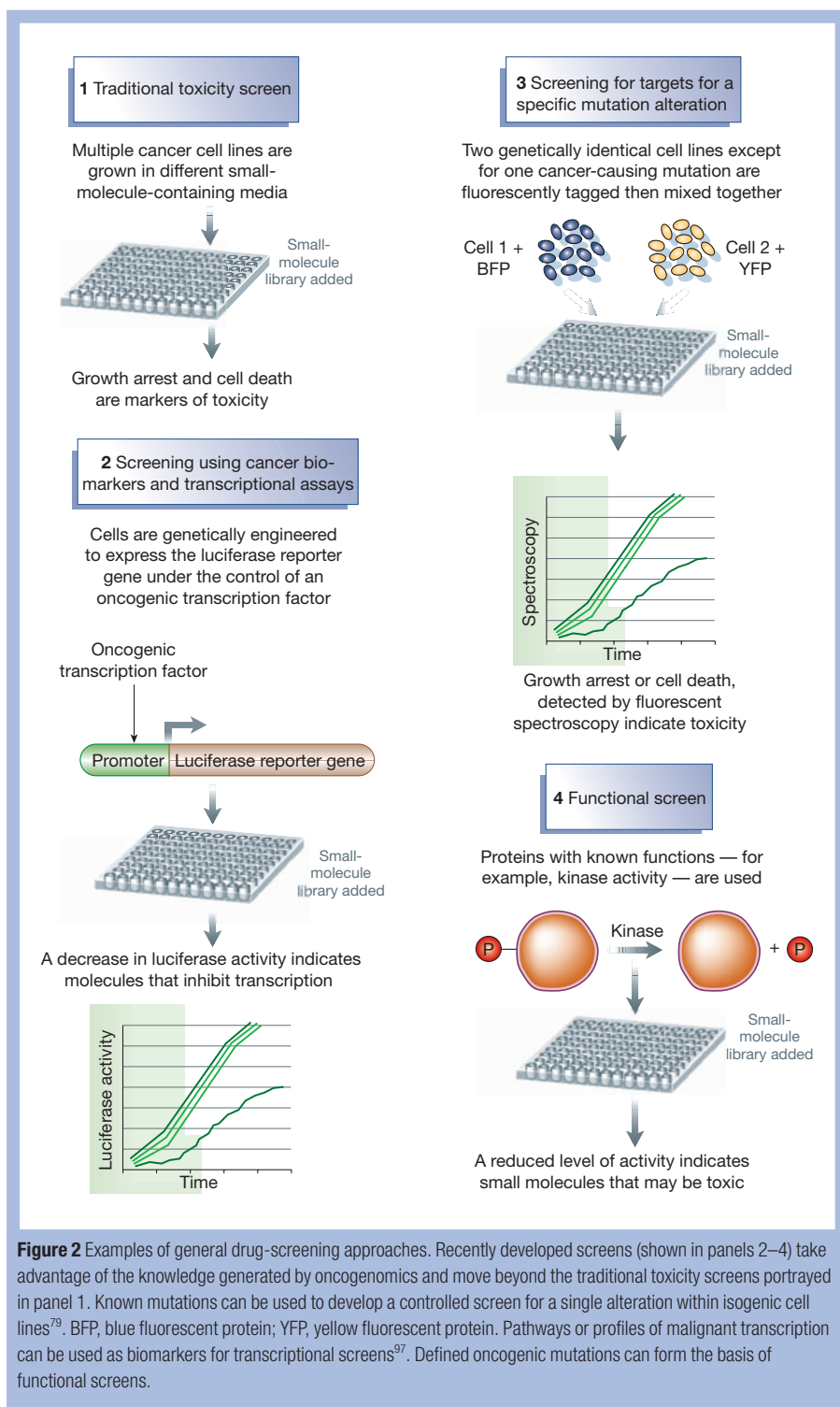


Figure 2 Examples of general drug-screening approaches. Recently developed screens (shown in panels 2–4) take advantage of the knowledge generated by oncogenomics and move beyond the traditional toxicity screens portrayed in panel 1. Known mutations can be used to develop a controlled screen for a single alteration within isogenic cell lines⁷⁹. BFP, blue fluorescent protein; YFP, yellow fluorescent protein. Pathways or profiles of malignant transcription can be used as biomarkers for transcriptional screens⁹⁷. Defined oncogenic mutations can form the basis of functional screens.

For much of the academic research community, the barriers preventing the development of immunotherapy reagents are formidable. However, they are not insurmountable. For example, the Ludwig Institute for Cancer Research has taken several potentially therapeutic antibodies, discovered by its own scientists, into clinical trials using reagents produced either in collaboration with industry or within its own biological production facilities^{91–93}. Moreover, numerous vaccine trials have been undertaken by the Institute on the basis of two important and widely expressed antigens, MAGE and NY-ESO-1, both of which were first discovered by scientists at the institute^{94,95}. Very recently, St Jude Children's Research Hospital announced the construction of a large in-house production facility

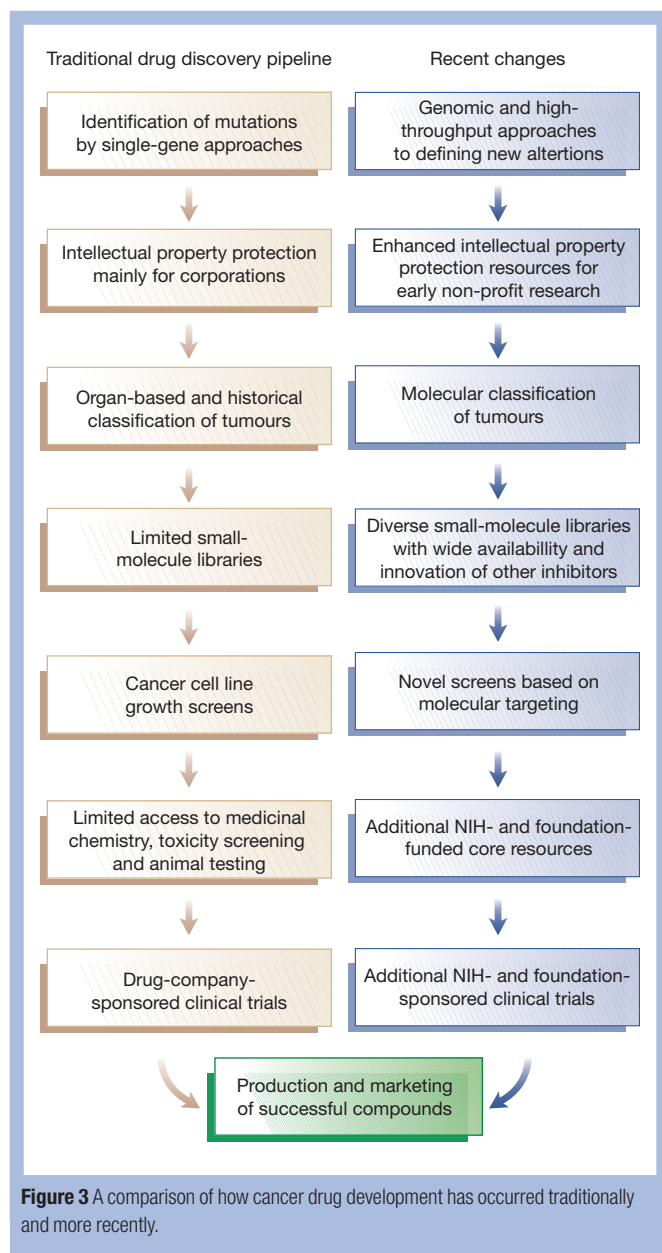


Figure 3 A comparison of how cancer drug development has occurred traditionally and more recently.

for generating novel therapeutic agents, and its own staff will take forward any leads into clinical trials.

Involving the entire research community

Over the past few decades, the complexity and difficulty in developing cancer interventions has become increasingly apparent. One informative indication of this challenge is the recent estimate that the total cost to develop a novel therapeutic is over US\$800 million — and much of this figure is accounted for by the number of failed leads⁹⁶. It is hoped that increased industry–academia collaboration in the early assessment of potential therapeutics will significantly reduce the number of failures, and thus improve the rate of development and bring down costs. Among the important advances of the HGP is a change in thinking about research organization based on the establishment of international teams of academic and industrial scientists. We think that such collaborations will be vital for the generation of cancer therapeutics. For this to occur, academic researchers will need to adopt a more prominent role in the leadership of early-stage clinical trials (Fig. 3).

We are at a crossroads where the community as a whole is accepting the responsibility for taking forward the advances that have been

made in basic genomics, not only into their focused application in identifying new therapeutic targets, but also in the development of new therapies. The range of talent, approaches and organizational structures that this will bring together will certainly increase the probability of significant successes and justify the hopes that launched the HGP when it was first initiated some 15 years ago. Indeed, it is only when we can convincingly argue that the cancer victim is the real beneficiary of this endeavour that success can be claimed. □

doi:10.1038/nature02627

- Vogelstein, B. & Kinzler, K. W. *The Genetic Basis of Human Cancer* (McGraw-Hill, New York, 2002).
- Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nature Genet.* **33** (suppl.), 238–244 (2003).
- Popescu, N. C. Comprehensive genetic analysis of cancer cells. *J. Cell. Mol. Med.* **4**, 151–163 (2000).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Druker, B. J. STI571 (Gleevec) as a paradigm for cancer therapy. *Trends Mol. Med.* **8**, S14–S18 (2002).
- Kim, J. A. Targeted therapies for the treatment of cancer. *Am. J. Surg.* **186**, 264–268 (2003).
- Smith, I. E. Efficacy and safety of Herceptin in women with metastatic breast cancer: results from pivotal clinical studies. *Anticancer Drugs* (suppl. 4), S3–S10 (2001).
- Salgaller, M. Technology evaluation: bevacizumab, Genentech/Roche. *Curr. Opin. Mol. Ther.* **5**, 657–667 (2003).
- Schadt, E. E., Monks, S. A. & Friend, S. H. A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem. Soc. Trans.* **31**, 437–443 (2003).
- Fischer, O. M., Streit, S., Hart, S. & Ullrich, A. Beyond Herceptin and Gleevec. *Curr. Opin. Chem. Biol.* **7**, 490–495 (2003).
- Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
- Bardelli, A. *et al.* Mutational analysis of the tyrosine kinase in colorectal cancers. *Science* **300**, 949 (2003).
- Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
- Hilgenfeld, E. *et al.* Spectral karyotyping in cancer cytogenetics. *Methods Mol. Med.* **68**, 29–44 (2002).
- Lucito, R. *et al.* Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**, 2291–2305 (2003).
- Cowell, J. K. & Nowak, N. J. High-resolution analysis of genetic events in cancer cells using bacterial artificial chromosome arrays and comparative genome hybridization. *Adv. Cancer Res.* **90**, 91–125 (2003).
- Liang, G. *et al.* DNA methylation differences associated with tumor tissues identified by genome scanning analysis. *Genomics* **53**, 260–268 (1998).
- Cottrell, S. E. *et al.* A real-time PCR assay for DNA-methylation using methylation-specific blockers. *Nucleic Acids Res.* **32**, e10 (2004).
- Greshock, J. *et al.* 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Res.* **14**, 179–187 (2004).
- Wang, T. L. *et al.* Digital karyotyping. *Proc. Natl Acad. Sci. USA* **99**, 16156–16161 (2002).
- Wang, T. L. *et al.* Digital karyotyping identifies thymidylate synthase amplification as a mechanism of resistance to 5-fluorouracil in metastatic colorectal cancer patients. *Proc. Natl Acad. Sci. USA* **101**, 3089–3094 (2004).
- Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA* **100**, 7696–7701 (2003).
- Beroud, C. & Soussi, T. The UMD-p53 database: New mutations and analysis tools. *Hum. Mutat.* **21**, 176–181 (2003).
- Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
- Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nature Genet.* **33**, 49–54 (2003).
- Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
- Rosenwald, A. *et al.* Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma. *J. Exp. Med.* **198**, 851–862 (2003).
- van 't Veer, L. J. *et al.* Expression profiling predicts outcome in breast cancer. *Breast Cancer Res.* **5**, 57–58 (2003).
- van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
- Nutt, C. L. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* **63**, 1602–1607 (2003).
- Dyrskjot, L. *et al.* Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genet.* **33**, 90–96 (2003).
- Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene-expression. *Science* **270**, 484–487 (1995).
- Lal, A. *et al.* A public database for gene expression in human cancers. *Cancer Res.* **59**, 5403–5407 (1999).
- Buckhaults, P. *et al.* Identifying tumor origin using a gene expression-based classification map. *Cancer Res.* **63**, 4144–4149 (2003).
- Jongeneel, C. V. *et al.* Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl Acad. Sci. USA* **100**, 4702–4705 (2003).
- Brentani, H. *et al.* The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl Acad. Sci. USA* **100**, 13418–13423 (2003).
- Strausberg, R. L. *et al.* An international database and integrated analysis tools for the study of cancer gene expression. *Pharmacogenomics J.* **2**, 156–164 (2002).
- Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).

40. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genet.* **36**, 40–45 (2004).
41. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99**, 16899–16903 (2002).
42. Xu, Q. & Lee, C. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.* **31**, 5635–5643 (2003).
43. Wang, Z. *et al.* Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.* **63**, 655–657 (2003).
44. Kriventseva, E. V. *et al.* Increase of functional diversity by alternative splicing. *Trends Genet.* **19**, 124–128 (2003).
45. Strausberg, R. L., Simpson, A. J. & Wooster, R. Sequence-based cancer genomics: progress, lessons and opportunities. *Nature Rev. Genet.* **4**, 409–418 (2003).
46. Dreves, J., Medinger, M., Schmidt-Gersbach, C., Weber, R. & Unger, C. Receptor tyrosine kinases: the main targets for new anticancer therapy. *Curr. Drug Targets* **4**, 113–121 (2003).
47. Sausville, E. A., Elsayed, Y., Monga, M. & Kim, G. Signal transduction — directed cancer treatments. *Annu. Rev. Pharmacol. Toxicol.* **43**, 199–231 (2003).
48. Cockerill, G. S. & Lackey, K. E. Small molecule inhibitors of the class 1 receptor tyrosine kinase family. *Curr. Top. Med. Chem.* **2**, 1001–1010 (2002).
49. Wilhelm, S. & Chien, D. S. BAY 43-9006: preclinical data. *Curr. Pharm. Des.* **8**, 2255–2257 (2002).
50. Joensuu, H. & Dimitrijevic, S. Tyrosine kinase inhibitor imatinib (STI571) as an anticancer agent for solid tumours. *Ann. Med.* **33**, 451–455 (2001).
51. Demetri, G. D. Targeting the molecular pathophysiology of gastrointestinal stromal tumors with imatinib. Mechanisms, successes, and challenges to rational drug development. *Hematol. Oncol. Clin. North Am.* **16**, 1115–1124 (2002).
52. Druker, B. Imatinib mesylate in the treatment of chronic myeloid leukaemia. *Expert Opin. Pharmacother.* **4**, 963–971 (2003).
53. Tipping, A. J. & Melo, J. V. Imatinib mesylate in combination with other chemotherapeutic drugs: *In vitro* studies. *Semin. Hematol.* **40**, 83–91 (2003).
54. Druker, B. Imatinib alone and in combination for chronic myeloid leukemia. *Semin. Hematol.* **40**, 50–58 (2003).
55. Joensuu, H. *et al.* Management of malignant gastrointestinal stromal tumours. *Lancet Oncol.* **3**, 655–664 (2002).
56. Joensuu, H. Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *N. Engl. J. Med.* **344**, 1052–1056 (2001).
57. GIST SU11248 Study Group. Clinical activity and tolerability of the multi-targeted tyrosine kinase inhibitor SU11248 in patients (pts) with metastatic gastrointestinal stromal tumor (GIST) refractory to imatinib mesylate. *Proc. Am. Soc. Clin. Oncol.* **22**, 814 (abstr. 3273) (2003).
58. Egland, K. A., Vincent, J. J., Strausberg, R., Lee, B. & Pastan, I. Discovery of the breast cancer gene *BASE* using a molecular approach to enrich for genes encoding membrane and secreted proteins. *Proc. Natl Acad. Sci. USA* **100**, 1099–1104 (2003).
59. Olsson, P., Motegi, A., Bera, T. K., Lee, B. & Pastan, I. *PRAC2*: a new gene expressed in human prostate and prostate cancer. *Prostate* **56**, 123–130 (2003).
60. Jager, D. Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. *Cancer Res.* **61**, 2055–2061 (2001).
61. Bera, T. K. *NGEP*, a gene encoding a membrane protein detected only in prostate cancer and normal prostate. *Proc. Natl Acad. Sci. USA* **101**, 3059–3064 (2004).
62. Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J. & Chen, Y. T. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol. Rev.* **188**, 22–32 (2002).
63. Scanlan, M. J., Simpson, A. J. & Old, L. J. The cancer/testis genes: review, standardization, and commentary. *Cancer Immunol.* **4**, 1 (2004).
64. Scanlan, M. J. *et al.* Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int. J. Cancer* **98**, 485–492 (2002).
65. Old, L. J. Cancer/testis (CT) antigens — a new link between gametogenesis and cancer. *Cancer Immunol.* **1**, 1 (2001).
66. Durrant, L. G. & Spendlove, I. Cancer vaccines entering Phase III clinical trials. *Expert Opin. Emerg. Drugs* **8**, 489–500 (2003).
67. Jager, E., Jager, D. & Knuth, A. Antigen-specific immunotherapy and cancer vaccines. *Int. J. Cancer* **106**, 817–820 (2003).
68. Albanell, J., Codony, J., Rovira, A., Mellado, B. & Gascon, P. Mechanism of action of anti-HER2 monoclonal antibodies: scientific update on trastuzumab and 2C4. *Adv. Exp. Med. Biol.* **532**, 253–268 (2003).
69. Blackledge, G. & Averbush, S. Gefitinib ('Iressa', ZD1839) and new epidermal growth factor receptor inhibitors. *Br. J. Cancer* **90**, 566–572 (2004).
70. Coiffier, B. Immunochemotherapy: the new standard in aggressive non-Hodgkin's lymphoma in the elderly. *Semin. Oncol.* **30**, 21–27 (2003).
71. Wannesson, L. & Ghilmini, M. Overview of antibody therapy in B-cell non-Hodgkin's lymphoma. *Clin. Lymphoma* **4** (suppl. 1), S5–S12 (2003).
72. Jain, R. K. Tumor angiogenesis and accessibility: role of vascular endothelial growth factor. *Semin. Oncol.* **29**, 3–9 (2002).
73. Milenic, D. E. & Brechbiel, M. W. Targeting of radio-isotopes for cancer therapy. *Cancer Biol. Ther.* **3** (2004).
74. Damle, N. K. & Frost, P. Antibody-targeted chemotherapy with immunoconjugates of calicheamicin. *Curr. Opin. Pharmacol.* **3**, 386–390 (2003).
75. McCormick, F. Cancer-specific viruses and the development of ONYX-015. *Cancer Biol. Ther.* **2** (suppl. 1), s157–s160 (2003).
76. Tong, A. W., Zhang, Y. A., Cunningham, C., Maples, P. & Nemunaitis, J. Potential clinical application of antioncogene ribozymes for human lung cancer. *Clin. Lung Cancer* **2**, 220–226 (2001).
77. Moon, C., Oh, Y. & Roth, J. A. Current status of gene therapy for lung cancer and head and neck cancer. *Clin. Cancer Res.* **9**, 5055–5067 (2003).
78. McNeish, I. A., Bell, S. J. & Lemoine, N. R. Gene therapy progress and prospects: cancer gene therapy using tumour suppressor genes. *Gene Ther.* **11**, 497–503 (2004).
79. Torrance, C. J., Agrawal, V., Vogelstein, B. & Kinzler, K. W. Use of isogenic human cancer cells for high-throughput screening and drug discovery. *Nature Biotechnol.* **19**, 940–945 (2001).
80. Strausberg, R. L. & Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **300**, 294–295 (2003).
81. Koehler, A. N., Shamji, A. F. & Schreiber, S. L. Discovery of an inhibitor of a transcription factor using small molecule microarrays and diversity-oriented synthesis. *J. Am. Chem. Soc.* **125**, 8420–8421 (2003).
82. Zerhouni, E. Medicine. The NIH Roadmap. *Science* **302**, 63–72 (2003).
83. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286–290 (2003).
84. Newman, D. J., Cragg, G. M. & Snader, K. M. Natural products as sources of new drugs over the period 1981–2002. *J. Nat. Prod.* **66**, 1022–1037 (2003).
85. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
86. Brizuela, L., Richardson, A., Marsischky, G. & Labaer, J. The FLEXGene repository: Exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era. *Arch. Med. Res.* **33**, 318–324 (2002).
87. Kretzschmar, T. & von Ruden, T. Antibody discovery: phage display. *Curr. Opin. Biotechnol.* **13**, 598–602 (2002).
88. Brekke, O. H. & Loset, G. A. New technologies in therapeutic antibody development. *Curr. Opin. Pharmacol.* **3**, 544–550 (2003).
89. Pasqualini, R. & Ruoslahti, E. Organ targeting *in vivo* using phage display peptide libraries. *Nature* **380**, 364–366 (1996).
90. Ellerby, H. M. *et al.* Anti-cancer activity of targeted pro-apoptotic peptides. *Nature Med.* **5**, 1032–1038 (1999).
91. Scott, A. M. *et al.* A Phase I dose-escalation study of sibrutuzumab in patients with advanced or metastatic fibroblast activation protein-positive cancer. *Clin. Cancer Res.* **9**, 1639–1647 (2003).
92. Welt, S. *et al.* Phase I study of anticolon cancer humanized antibody A33. *Clin. Cancer Res.* **9**, 1338–1346 (2003).
93. Scott, A. M. *et al.* Specific targeting, biodistribution, and lack of immunogenicity of chimeric anti-GD3 monoclonal antibody KM871 in patients with metastatic melanoma: results of a phase I trial. *J. Clin. Oncol.* **19**, 3967–3987 (2001).
94. Jager, E. *et al.* Induction of primary NY-ESO-1 immunity: CD8+ T lymphocyte and antibody responses in peptide-vaccinated patients with NY-ESO-1+ cancers. *Proc. Natl Acad. Sci. USA* **97**, 12198–12203 (2000).
95. Atanackovic, D. Vaccine-induced CD4+ T cell responses to MAGE-3 protein in lung cancer patients. *J. Immunol.* **172**, 3289–3296 (2004).
96. Duyk, G. Attrition and translation. *Science* **302**, 603–605 (2003).
97. Rapisarda, A. *et al.* Identification of small molecule inhibitors of hypoxia-inducible factor 1 transcriptional activation pathway. *Cancer Res.* **62**, 4316–4324 (2002).

Acknowledgements We thank G. Demetri for discussions and advice during the preparation of this manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

The case for a US prospective cohort study of genes and environment

Francis S. Collins

National Human Genome Research Institute, National Institutes of Health, Building 31, Room 4B09, MSC 2152, 31 Center Drive, Bethesda, Maryland 20892-2152, USA (e-mail: fc23a@nih.gov)

Information from the Human Genome Project will be vital for defining the genetic and environmental factors that contribute to health and disease. Well-designed case-control studies of people with and without a particular disease are essential for this, but rigorous and unbiased conclusions about the causes of diseases and their population-wide impact will require a representative population to be monitored over time (a prospective cohort study). The time is right for the United States to consider such a project.

Identification of the genetic and environmental factors that contribute to health, disease and response to treatment is essential for the reduction of illness. This, of course, is the primary goal of biomedical research. Several auspicious recent developments suggest that progress in this area could be quite rapid. The sequence of the human genome^{1,2} and increasing information about the genome's function have provided a robust foundation for the investigation of human health and disease. Likewise, results from the exploration of human genetic variation through the International HapMap Project³ will soon furnish researchers with a powerful tool for identifying variants that contribute to common disease. This information will be especially useful when combined with reliable, cost-effective, high-throughput methods that can be used to genotype these variants in large population samples⁴.

In parallel with the expansion of genomic tools and knowledge, methods for measuring non-genetic factors and

environmental exposure have improved. These techniques promise to extend the range of epidemiological investigation⁵. There is growing recognition that a change in the environment, in combination with genetic disposition, has produced most recent epidemics of chronic disease, and may hold the key for reversing the course of some diseases⁶. For example, consider the interaction of presumed famine-protective genetic predispositions with a modern environment in which there is a ready availability of excess calories. This has probably contributed to the current obesity epidemic in the United States. Development of robust analytical methods for assessing disease-risk relationships and interactions is beginning to allow researchers to disentangle such complex effects on a population scale⁷.

Together, these developments present an exciting opportunity to address unanswered questions related to the complex contributions of genes, the environment, and gene-gene and gene-environment interactions to



Rigorous quantitative assessment of genetic and environmental risk factors will be critical for the future of medicine.

Box 1

Desirable characteristics of a gene–environment cohort study

To maximize the value of a prospective cohort study for determining gene, environment, gene–gene and gene–environment contributions to common disease, it should have many, if not all, of the following characteristics.

- A large number of participants, at least several hundred thousand, should be enrolled. This would ensure an adequate sample size for common disorders, particularly for gene–environment interactions.
- Minority groups should be intentionally over-sampled to permit meaningful inferences about these groups and for the study of health disparities.
- A broad range of ages should be represented to provide information on disorders from infancy to old age, with over-sampling of age groups as needed.
- A broad range of genetic backgrounds and environmental exposures should be included to provide enough variability to detect and compare associations and interactions.
- Family-based recruitment, including multiple generations, should be used for at least part of the cohort to increase the power of genetic analyses.
- A broad array of clinical and laboratory information, not limited to any single disease, should be collected at the beginning and at regular intervals thereafter.
- Sophisticated dietary, lifestyle and environmental exposure assessments should be carried out, using both questionnaires and biological measures.
- Biological specimens, including DNA, plasma and cells, should be collected and stored.
- A highly sophisticated data-management system should be included.
- Access to study data and biological materials should be free and open to allow research into many diseases by scientists in many sectors.
- Investigations during the study should not be limited to hypotheses conceived at its inception.
- Comprehensive community engagement should be a major feature in the design and implementation of the study.
- A state-of-the-art consent process should be adopted to allow multiple uses of the data and regular feedback to participants about progress.

health. Understanding these factors and their interactions could lead to major improvement in diagnostics, preventive medicine and therapeutics.

Case–control studies and beyond

A widely used and highly successful approach to identifying factors that contribute to specific illnesses is the case–control study. For this, carefully chosen people with and without a disease are analysed for differences in the distributions of genetic variation and/or environmental exposures or other non-genetic factors⁸. Valuable insights, perhaps unobtainable in any other way, have been derived from such studies, particularly for rare disorders. False-positive genetic associations related to differences among various population groups have been a problem in the past, but the availability of high-throughput, low-cost genotyping can reduce this risk by pre-matching genetic markers with a panel of random ones or otherwise adjusting for background genetic differences⁹.

Case–control studies have certain weaknesses, including the tendency for clinically diagnosed cases to represent the more severe end of the disease spectrum¹⁰ and the difficulty of selecting an unbiased control group. In addition, many case–control studies are plagued by the problem of recall bias: memories of individuals diagnosed with disease are often coloured by their subsequent experience of illness¹¹. For example, a recent case–control study of coronary heart disease showed that people with heart disease were more likely to report a family history of the disease that could not be verified than were controls¹². Although this problem may not affect genotype-specific risk-ratio calculation¹³, it is still a significant problem for the overall assessment of disease risks.

So, although the case–control strategy can be a powerful means for identifying potential risk factors, its inherent biases make the quantification and population-wide generalization of risk difficult. Replication of associations and estimation of their magnitude, consistency and temporality (all key criteria for epidemiological evidence of causal relationships¹⁴) are best obtained through prospective, population-based cohort studies⁸.

To appreciate the contrasting but potentially complementary nature of case–control and prospective cohort studies, consider the example of diabetes. A case–control study of 5,000 cases and 5,000 controls could be mounted over a year or two, and could be used to identify susceptibility genes and environmental correlates of risk. But selection biases for the phenotype (for instance, a previously known

diagnosis of diabetes) would prevent quantitative generalization of the results. Furthermore, such a study would probably be subject to recall biases among the cases about their family history, diet and other environmental factors, and there would be no specimens available from the people being studied before diagnosis to search for predictive biomarkers. These shortcomings could be addressed by a longitudinal study of 200,000 people, but it would probably take several years for 5,000 of them to develop diabetes. In the long run, however, the need for this kind of information for 40 common diseases would require the collection of data on 200,000 people anyway, and the prospective cohort study would also allow links to other conditions (such as hypertension and obesity) to be detected.

Identifying logistical hurdles

Along with the many advantages of prospective studies is a unique set of challenges, most of which centre on logistics. Such studies generally require large sample sizes, detailed characterization at the beginning of the study, and prolonged follow-up for the occurrence of most common chronic diseases⁸ (see Box 1).

Large-scale cohort studies are under discussion or already underway in the United Kingdom (the UK BioBank), Iceland (deCODE), Estonia, Germany, Canada and Japan. Although such projects are likely to be useful for research everywhere, the United States should seriously consider undertaking a national investigation of its own. Inadequate representation of important US minority groups who bear disproportionate burdens of disease (particularly African-Americans, Latinos and Native Americans), the probable presence of different environmental risk factors, and the potential for limited access to data and biological materials make it unlikely that the current cohort projects will be adequate for the needs of the United States.

In the United States, a gene–environment cohort study could be assembled by building, at least in part, on already existing large studies such as the Women's Health Initiative, the Framingham Study, the Harvard studies of health professionals, and some of the many large cancer cohorts. The obvious advantages are that many years of follow-up have already taken place in these cohorts and, for many of them, DNA has already been collected. But serious consideration must be given to whether the disease-specific focus of many of these studies has limited the phenotyping and exposure measures, whether the minority representation is adequate, whether the consent obtained is sufficient for broad access to data and biological materials, and whether the study design is appropriate for the

ambitious goals of a national gene–environment study. If those limitations turn out to be significant, an entirely new cohort project may need to be contemplated.

Evaluating the merits

Although the challenges in undertaking such a prospective population study in the United States will be considerable, a serious evaluation of its merits is now in order. This debate should engage a wide variety of experts in epidemiology, genetics, environmental science, ethics, public health, economics and public policy. An initial meeting at the National Institutes of Health in December 2003 led to agreement that such an effort should be explored further. If the conclusion is that this resource is needed, then we must collectively seek ways to organize and implement it quickly and efficiently — or face the real possibility that a decade from now the promise of genetic and environmental research for reducing disease burden on a population basis will remain out of reach. □

doi:10.1038/nature02628

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
4. Shi, M. M. Technologies for individual genotyping: detection of genetic polymorphisms in drug

- targets and disease genes. *Am. J. Pharmacogenomics* **2**, 197–205 (2002).
5. Weaver, V. M., Buckley, T. J. & Groopman, J. D. Approaches to environmental exposure assessment in children. *Environ. Health Perspect.* **106** (suppl. 3), 827–832 (1998).
6. Chakravarti, A. & Little, P. Nature, nurture and human disease. *Nature* **421**, 412–414 (2003).
7. van den Oord, E. Method to detect genotype–environment interactions for quantitative trait loci in association studies. *Am. J. Epidemiol.* **150**, 1179–1187 (1999).
8. Gordis, L. *Epidemiology* 2nd edn (W. B. Saunders & Co., Philadelphia, 2000).
9. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
10. Guo, S. W. Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. *Am. J. Hum. Genet.* **63**, 252–258 (1998).
11. Feinstein, A. R. Experimental requirements and scientific principles in case–control studies. *J. Chronic Dis.* **38**, 127–133 (1985).
12. Silberberg, J. S., Włodarczyk, J., Fryer, J., Ray, C. D. & Hensley, M. J. Correction for biases in a population-based study of family history and coronary heart disease. The Newcastle Family History Study I. *Am. J. Epidemiol.* **147**, 1123–1132 (1998).
13. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356–1360 (2001).
14. Hill, A. B. The environment and disease: association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).

Acknowledgements Appreciation is expressed for comments on the manuscript from J. Bailey-Wilson, W. Burke, M. Boehnke, B. Foxman, A. Guttmacher, E. Lander, T. Manolio, A. Wilson and E. Zerhouni. The opinions expressed in this Commentary are those of the author, and do not represent an official position of the National Human Genome Research Institute, the National Institutes of Health or the Department of Health and Human Services.

Competing interests statement The author declares that he has no competing financial interests.

Organizational challenges in clinical genomic research

Jill S. Altshuler¹ & David Altshuler²

¹AltshulerGray LLC, 61 Dean Road, Brookline, Massachusetts 02445, USA (e-mail: jill@altshulergray.com)

²Departments of Genetics and Medicine, Harvard Medical School and Massachusetts General Hospital, and Broad Institute, 55 Fruit Street, Wellman 8, Boston, Massachusetts 02114, USA (e-mail: altshuler@molbio.mgh.harvard.edu)

Genome sequence data are enabling clinical genomic investigation, in which the characteristics of human patients are explored using comprehensive inventories of biomolecules. Successful investigators must navigate rapid technological change, collect and analyse large volumes of data, and engage systems of clinical care. Such projects will increasingly rely on fully integrated multidisciplinary teams, demanding new organizational models in academic biomedical research.

Until recently, genome science and human clinical research were independent fields, each with distinct intellectual traditions and communities of investigators. With the completion of the Human Genome Project, there is great interest in combining these disciplines to study disease mechanisms and improve patient care, an endeavour that we term 'clinical genomic investigation'. Whereas other reviews in this issue consider scientific approaches to this goal, we discuss the substantial organizational challenges inherent in combining these two fields.

The challenges

The field of genomics is built on two core capabilities: collecting comprehensive data sets using technologically advanced laboratory tools, and searching for subtle relationships within them using computationally sophisticated analytical methods. Previously, neither was required for success in biomedicine, and yet both are now indispensable for investigation aiming to connect genomic information with human disease.

Genomics has experienced a rapid pace of technological change, requiring that scientists keep up with emerging technologies, and skillfully integrate complex laboratory and information systems. The potential advantages of each new technology must be balanced against a typically short half-life and substantial cost, and this calculation endows the management of technology with considerable strategic importance. The process of collecting, storing and distributing large data sets places an emphasis on management skill and software engineering. Computational analysis of these large data sets draws on expertise from mathematics, statistics and computer science. Although computational scientists have long been valued in fields as diverse as physics, finance and telecommunications, they have been relatively few in biomedicine.

Research involving human subjects poses a unique set of challenges, largely attributable to the complexity of the human population, the need to respect patient autonomy, and pressure on highly distributed systems of clinical care. Human studies require careful attention to informed consent, and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) codifies even greater attentiveness to privacy and security of individual patient information. The demands of informed consent and data protection compete with financial pressures that have left clinicians barely enough time for patient care, let alone research. The local

jurisdiction of institutional review boards, and the lack of universal information technologies for managing patient information (much of which is still on paper, or in legacy computer systems) make it difficult to compare data across institutions even after HIPAA regulations have been met.

If both genomics and clinical investigation are complex in isolation, combining them multiplies the challenge. Human genetic research places greater demands on privacy protection, because a single DNA sample contains a complete set of genes, and thus could be misused to study hypotheses beyond the initial consent. Variations in DNA sequence provide a unique DNA 'fingerprint', which in theory could be used to identify individual study participants, or related to societal concepts of race, ethnicity and group identity. Genomic studies of human populations require larger sample sizes than do genomic investigations of well-controlled model systems, because variation in genotypic background, environment and behaviour all introduce noise. This is particularly problematic in unbiased genome-wide studies that consider 25,000 or so genes, each of which has a correspondingly low chance of being involved in the process of interest (see review in this issue by Carlson, page 446). The increased statistical significance required to distinguish the signal from the noise can be achieved by increasing the sample size. But collecting large samples requires many investigators and coordination across multiple institutions, with all the logistical challenges mentioned above.

Organizational implications

Most biomedical research is well served by the traditional model of individual laboratories led by a single principal investigator, but clinical genomic research often requires multiple investigators. This has focused attention on the importance of teamwork in biomedical science. (It is a core feature of the recently announced NIH Roadmap; see <http://nihroadmap.nih.gov> and Box 1.) But teams can be constructed in many different ways, each with its own capabilities and requirements for success (see Box 2). We think that clinical genomic research will most successfully be conducted by mission-focused, interdisciplinary teams that are fully integrated (rather than virtual) and work together over a sustained period to solve specified research questions.

A review of large-scale multidisciplinary research efforts suggests three common characteristics of successful, integrated teams. Chief among these is a clear and compelling



Attention to teamwork has secured wins for the New England Patriots.

mission — a grand challenge that attracts, motivates and unites a set of disparate researchers and inspires them to subjugate their individual needs and egos in the interest of a common goal. The mission must be larger than that achievable by an individual, and yet narrow enough to have specific meaning. Crafted well, a mission provides a subtle form of guidance for consistent decision making across a distributed leadership, minimizing the need for a rigid hierarchical structure that would limit creativity and motivation.

For example, in the 1970s, Xerox PARC in Palo Alto, California, articulated its mission: to create the 'office of the future'. Attracting a body of like-minded researchers from fields as diverse as organizational behaviour, computer science and physics, PARC created the computer mouse, the 'what you see is what you get' (WYSIWYG) computer environment, personal computers and the laser printer. An example from within biomedicine is the Human Genome Project, in which a large number of investigators from different disciplines (biology, mathematics and engineering) together tackled a daunting scientific challenge, created new technologies and organizational structures, shared data to an unprecedented degree, and as the project matured, analysed and published the results collaboratively. Other instances include the synthesis of bioactive steroid hormones (discussed, along with other examples, in ref. 1) and the development of protease inhibitors to treat HIV².

The second requirement for success is the selection of team members — people who commit to the success of the group even at the expense of individual achievement. The field of high-energy physics has institutionalized this notion by requiring — and placing high value on — service roles. At Fermilab (Batavia, Illinois), for instance, all physicists commit a substantial fraction of their time to fulfilling core support functions, such as creating and maintaining shared equipment, managing data collection, and providing computational support. In recruiting for such teams, disciplinary excellence is not the sole criterion: individuals must be flexible and open-minded, have good communication and social skills, and be willing to work with others in pursuit of a common goal. The National Football League's New England Patriots are a striking example of such a human resource strategy: they won the Super Bowl (league championship) twice in three years by jettisoning individual superstars, identifying undervalued players who would thrive in a particular role and system, and distributing credit over the entire team.

Third, leaders of multidisciplinary teams must be skilled at motivating and facilitating the work of others, inspiring people with different backgrounds and career goals without necessarily having formal authority over their careers. Given a potentially fragile mix of skills and perspectives, team leaders must guide individuals to contributions that advance the shared goal, manage communication and group dynamics, and promote the good of the team. This can be achieved through the rare combination of charisma and selflessness. As a physicist who had a high-ranking position at Bell Labs in the 1970s told one of us, the most effective leaders play "almost a service role", with an uncanny ability to lead despite a manner that communicates "nobody works for me; I work for them".

How well do these characteristics map onto the organizational model typically encountered in academic biomedical research? Unfortunately, not very well. The notion of a mission uniting multiple investigators may be viewed suspiciously by a community that celebrates academic freedom and creativity, as well as individual success. More importantly, the current incentive structure actively reinforces individualism over teamwork. In biomedicine, individuals tend to be rewarded rather than teams: successful faculty members enjoy tenure, space and resources to advance their ideas and substantial salaries, but team members (typically trainees and transient technical staff) have little in the way of status, control, job security or financial rewards. Systems for publication reinforce a 'winner takes all' mindset, allowing only two prominent roles on publications, that of first and senior authors. In multi-investigator groups, only a few contributors can be highlighted on any given paper.

Box 1

The NIH Roadmap

In September 2003, Elias Zerhouni, director of the National Institutes of Health (NIH) in Bethesda, Maryland, announced that the agency would begin a series of new initiatives that constitute a new 'NIH Roadmap'. The Roadmap initiatives are the product of a series of meetings Zerhouni initiated with scientists, academics, business leaders and members of the public after he was appointed NIH director in May 2002. Zerhouni says the plans lay the groundwork for the future of biomedical research, which will be conducted by teams of interdisciplinary researchers who will need new tools to transform basic discoveries into new treatments. The Roadmap aims to provide those tools and spur the development of new types of scientific team.

In its first year, the NIH will spend US\$130 million on Roadmap initiatives; by 2009, Zerhouni hopes that the NIH will have spent about US\$2.1 billion on the initiatives. Roadmap initiatives fall into three themes: New Pathways to Discovery, Research Teams of the Future, and Re-Engineering the Clinical Research Enterprise. The Roadmap initiatives focus mainly on the development of tools, technologies, networks and resources, instead of funding individual projects.

For instance, some of the first projects announced under the Roadmap were grants to fund extramural biomedical computing centres, small-molecule repositories and screening centres, and proteomics-technology development centres.

Another main focus of the Roadmap is on high-risk research, and one new project — the Director's Pioneer Award — will award up to ten grants of half a million dollars each for five years to individuals who have "exceptionally creative abilities and diligence", according to the award announcement.

Finally, the Roadmap envisions a revamping of the nation's clinical research infrastructure by linking doctors in private offices with community and patient groups and large health care networks. The Roadmap would create a group of new clinical research associates, in an effort to establish an infrastructure that can be used for many different clinical studies.

More information on the Roadmap, including funding opportunities, can be found at <http://nihroadmap.nih.gov>.

Erika Check

Box 2

Models of teamwork in science

When the demands of a research question exceed the ability of a single lab to pursue them, there are at least three strategies for obtaining additional expertise and capabilities, each of which can be construed as a team-based approach: outsourcing, insourcing and integration.

Outsourcing

The simplest (and traditional) approach is to increase the scale and disciplinary focus of individual labs by collaborating with other research groups. This model deliberately *outsources* such capabilities and requires little organizational change, but reaches its limitations with larger-scale, sustained projects. Large, interdisciplinary projects require coordination, which can be difficult to achieve if each lab and participant is guided by personal interest (and rewarded according to individual accomplishment), rather than by an explicit and shared goal. Substantial investment of time and effort may be required before rewards (for example, results and publications) can be realized, a barrier that will not be overcome without commitment to a sustained and defined relationship among collaborators.

Insourcing

A second model retains the disciplinary focus of labs led by individual principal investigators, but brings new capabilities in-house through 'core labs'. This can be viewed as a decision to *insource* an otherwise unavailable, expensive and technically sophisticated capability. Core labs involve the creation of discrete entities whose goal is to assist other researchers (rather than to pursue specific questions), and are most often run by staff scientists, rather than by faculty and trainees. Valuable in the right setting, core labs are generally most successful when the service they offer is a commodity — well defined and

technologically mature. In cases such as genomics, where technology is rapidly evolving and demand for practitioners is high, it can be difficult to recruit and retain the few scientists who have experience of establishing and running such capabilities — particularly to a purely service role. The idea of core labs for statistical support and computation is a particularly poor fit in the area of genomics, where analytical methods are nascent, often must be customized for each application, and represent one of the more intellectually sophisticated (and limiting) contributions to success.

Integration

A third model pays less attention to traditional barriers between individual labs and disciplines, instead creating an interdisciplinary team united in pursuit. Members of this team interact as partners rather than being under the sole leadership of a single faculty member or operating in a client-service model. Rather than isolating staff with specialized knowledge in core labs, separate from the scientific questions that motivate them, the goal here would be to fully *integrate* a range of capabilities and disciplines to solve a particular set of problems. There is a crucial element of scale, however. Such teams must be large enough to encompass the required range of capabilities, disciplines and efficiencies, yet small enough to remain focused on a single mission, communicate and respond to new information, and maintain a sense of shared purpose. Others, notably Brown and Goldstein, have written about the catalytic value of long-term collaborations that bring together like-minded investigators with discrete and complementary capabilities¹. The teams we envision are similar in spirit, if somewhat larger in scale and broader in disciplinary focus.

Institutional change

Allowing sustained, multidisciplinary teamwork to become a viable and attractive option for biomedical research demands cultural and structural change — change that will only be brought about by academic institutions and the faculty who work in them. We focus on two areas of particular importance: education and career development.

First, scientific training for clinical genomics should involve team-based learning and greater efforts should be made to break down disciplinary barriers. Learning in groups, as well as working together on specific projects, might help young doctors and scientists to become accustomed to teamwork — not to mention to the notion that they are part of one biomedical research culture, not two. Moreover, shared activities and language promote mutual respect and understanding, which is essential for effective, sustained collaboration. (Note that training specialists with distinct skills to work together is a different goal from that of training individuals such as MD/PhDs to have dual competency.) For example, in one Boston programme, biomedically oriented physics and engineering PhD students undertake clinical rotations: on a recent programme review, alumni felt that this clinical exposure was an essential element of the programme and had deeply influenced their careers⁵. Such a curriculum will be effective only if offered to a receptive student body. In addition to the usual measures of academic potential, admissions committees may want to evaluate communication skills, ability to work in teams, and the capacity to synthesize information across disciplines.

Second, to develop the careers of people who work in teams, institutions must implement improved methods to evaluate contributions to collective accomplishments. This must start with a thorough rethinking of authorship — how it is assigned and the role it has in the selection and promotion of scientists. One way to avoid the distortions of the current first author/last author model would be to emulate physics, listing authors alphabetically. Although this model has a certain egalitarian appeal, it seems to us a missed opportunity to provide

more, rather than less, information about individual contributions. (Instead it places an unrealistic burden on tenure committees to unearth the particular roles of the scientist under review, and may perhaps invite self-promotion.) We would rather see a deconstruction of the author list, where the particular contributions of each author are specified by the team itself. Such a system increases opportunities for recognition. For example, it would allow individual contributions to be acknowledged, be it patient sample collection, genomic technology, statistical analysis or biological follow-up. Each scientist or clinician would be highlighted for his or her disciplinary contributions. Consider the credits of a feature film: although the stars get top billing and the director comes last, the cinematographer and costume designer get their due. Such credits make it possible to evaluate and recognize specific contributions in the context of an ensemble project.

Evolution in the recording of authorship would be an important step towards revamping scientific career development. Committees that select and promote faculty staff often place weight on quantitative authorship-related measures (such as impact ratios and the number of papers published) and outside appraisal, but in future they will need to seek out and incorporate new assessments of scientific contribution for both their faculty and their non-faculty professional staff. But how can senior faculty, already stretched thin, be expected to take on this added responsibility? One answer may be to look at the way professional service firms work, such as law and consulting firms, where — similar to academia — people are the most valued assets, and career development is thought to be sufficiently important that senior professional staff directly oversee it themselves. In such firms, dedicated administrative staff devote substantial effort to designing procedures for evaluation, gathering input from co-workers and clients, and managing the process to ensure that the time of senior staff is efficiently used. Although the details may differ, it would appear wise to invest in better procedures for evaluating academic contributions, rather than allowing the

existing system, with its well-documented shortcomings, to dictate the meaning of success in universities^{3,4}.

A move towards team-based working raises a set of challenges that academic institutions will have to address. As well as providing a fair means of evaluating team members, managers need to ensure that teams don't succumb to 'groupthink', where ideas from outside the team are devalued or viewed suspiciously. Poorly managed teams can suffer from diffusion of responsibility and accountability. Even the most highly motivated and functional team may outlive its mission, which illustrates the need for mechanisms to disband the team and transfer personnel. Addressing these and other challenges will require attention, creativity and good management.

A key to success

Team-based science is not simply a fad, but a reasoned response to the fundamental challenges of combining clinical investigation with genome-wide hypothesis generation. Although many different

organizational models can and should be explored, we believe that a subset of clearly articulated and important problems can effectively be studied through sustained, goal-orientated, multidisciplinary teams. Creating and nurturing these teams will require open-mindedness and imagination in considering new approaches to recruitment, education, evaluation and rewarding of success in academia. □

doi:10.1038/nature02629

1. Goldstein, J. L. & Brown, M. S. The clinical investigator: bewitched, bothered, and bewildered — but still beloved. *J. Clin. Invest.* **99**, 2803–2812 (1997).
2. Galambos, L. & Sewell, J. E. *Confronting AIDS: Science and Business Cross a Unique Frontier* (Merck, Whitehouse Station, New Jersey, 1998).
3. Kennedy, D. Multiple authors, multiple problems. *Science* **301**, 733 (2003).
4. Editorial. Who'd want to work in a team? *Nature* **424**, 1 (2003).
5. Gray, M. L. & Bonventre, J. V. Training PhD researchers to translate science to clinical medicine: closing the gap from the other side. *Nature Med.* **8**, 433–436 (2002).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequinot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: r.asami@naturejp.com

naturejobs

Back to life

In a report aptly titled *Resurgence*, analysts Ernst & Young this month revealed that the global biotech sector is at last finding its feet. Although the past two years saw declining stock values and a sharp shortfall in venture capital for the area, the companies that survived these harsh conditions are now beginning to improve their performance. This is partly because these firms at last have products available on the market.

The new lease of life is already being reflected in recruitment — the sector saw some 20,000 new posts arise in 2003, a 9% increase on the previous year. And with many other products nearing the end of their development stage — 296 are currently undergoing phase III clinical trials — the outlook for the future seems to be bright.

Another indicator of the sector's increasing strength is the rise of 'big biotech' — a dozen or so large companies that are dominating the market. These have emerged from a series of mergers, such as the formation of Biogen Idec, which is worth some US\$6.8 billion and became the the third-largest biotech company in the United States when Biogen and IDEC Pharmaceuticals joined together last year. Such unions usually allow the partners to make up for each other's shortcomings — be they marketing, products or cash — and, once the dust has settled, they tend to result in a more stable working environment.

But scientists who are contemplating a career in biotechnology still need to pay attention to their prospective employer's bottom line (see page 484). In particular, they should check that the firm has sufficient funds to sustain itself for the next two years. This is important because among the publicly traded biotech companies, 31% in the United States, 43% in Europe and 56% in Canada are likely to run out of cash in that time. The lesson, as ever, is that the future comes with no guarantees — so beware.

Paul Smaglik

Naturejobs editor



Contents

SPECIAL REPORT

Taking a risk in start-ups p484

CAREER VIEW

Recruiters & Industry

Rules of engagement

Graduate Journal

Wrapping things up

Movers

Eve Slater

p486

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

If you are planning to start your career by joining a biotech start-up, you will need to do some homework.

Virginia Gewin assesses the risks involved.

Virologist John Alderete made a flying start when he co-founded the biotech start-up Xenotope Diagnostics in 2001. Within 18 months, he and chief executive Paul Castella had raised US\$1.2 million, developed a test for a sexually transmitted disease and had it approved by the US Food and Drug Administration (FDA). But this high-speed approach left the company, based in San Antonio, Texas, teetering on the cusp of financial disaster — it wasn't easy to navigate the regulatory, funding and clinical-trial hurdles to an ultimately successful licensing agreement with Genzyme Diagnostics. "The start-up world is not for the faint of heart," says Alderete.

Needless to say, choosing a job with a new biotech company is nearly as risky as starting one up. Industry leaders recommend basing your choice of employer on the same criteria used by venture capitalists to decide whether they will fund a start-up.

The hard truth is that great science doesn't necessarily meet the bottom line. "How much risk is somebody willing to take?" asks Michael Steinmetz, a partner in the life-science investment firm MPM Capital of Cambridge, Massachusetts. Only about a third of companies make it, meaning that they are sold

for profit or bring big returns to investors, he says.

In addition to taking a healthy dose of risk, scientists seeking employment in industry give up both intellectual freedom and bench-side competition in return for higher wages and collaboration. "Business is about team success, not personal success," says Caitlyn Waller of BayBio, a San Francisco-based bioscience industry organization.

For this reason, biotech leaders recommend that prospective employees assess their personality before making the leap from academia. "If you want to do cool research and be able to play, don't go into the biotech industry," says Carl Johan Sundberg, investment manager at the Karolinska Investment Fund in Stockholm, Sweden. Investor Steven Burrill, chief executive of San Francisco's Burrill and Company, agrees that prospective employees tend to forget that industry is about making money. "We finance businesses, not science projects," he says.

But an industry career no longer means having to relinquish one's scientific standing. Adeela Kamal, a researcher at Conformia Therapeutics in San Diego, California, doesn't feel that her scientific status has been compromised by her industry work. In fact, her high-profile work in novel cancer therapeutics has garnered invitations to speak at top conferences and write review articles.

Few industrial scientists spend their whole career with one company, so prospective employees should realize they are exchanging security for short-term experience. "It doesn't mean you shouldn't take a risk, just understand the risk you are taking," says Scott Morrison, US director of life sciences at financial analysts Ernst & Young in San Francisco.

DO YOUR HOMEWORK

Business leaders agree on the top three points to evaluate when considering a job at a start-up: the management team's experience, the amount of venture-capital funding and, of course, the science.

"You must give due diligence to the people and science behind the company," says Otello Stampacchia, managing partner at Biotechnology Private Equity Advisors in Geneva, Switzerland. Look for an

Human resources

From Stanford University's Bio-X programme to the California Institute for Quantitative Biomedical Research, a number of academic institutions are creating bridges between academia and industry. The Center for BioEntrepreneurship (CBE) at the University of California, San Francisco, provides networking opportunities for industry mentors, scientists and academic inventors. "The goal is to help the student and postdoc population to become leaders in the life-sciences industry," says the centre's director, Katherine Moortgat.

The CBE runs courses as well as providing start-up finance. Of the roughly 30 business plans and teams developed with support from the centre in the past three years, at least eight received significant venture-capitalist interest.

V.G.



Understanding risk: (clockwise from top left) Magnus Persson, Jesper Zeuthen, Steven Burrill and Katherine Moortgat recommend gaining practical experience before entering the world of start-up companies.

CORBIS

experienced management team — serial entrepreneurs with multiple successes under their belt, or those participating in regional industry organizations. Find out about less-well-known entrepreneurs' science by searching PubMed.

In addition, learn what companies the top venture capitalists are funding, and why. Waller suggests looking for patterns in local and national venture conferences. "The strong companies are picked to present at conferences," she says. "Use it as a pre-screening process."

Choosing a position involves making an informed financial decision. Don't be afraid to ask to see the company's financial status in an interview. If the company doesn't have solid funding for at least two to three years, you may want to reconsider.

ALTERNATIVE ROUTES

Remember that research isn't the only avenue for a science career. Employers need scientists in many other areas, including regulatory affairs, intellectual property and clinical design. Although you may need to broaden your scope of knowledge for many of these positions, any business background will be a boon.

"Practical experience is what is called for, not adding 'MBA' to your title," says Jesper Zeuthen, managing director of BankInvest Biomedical Venture in Copenhagen, Denmark. But you need to learn the language, he adds. At the most basic level, you have got to know your 'assets' from your 'business model'.

But networking remains key. "At the end of the day, a lot depends on the network that you manage to build," says Stampacchia, who got his position in part thanks to a conversation he started on a plane. He and other venture capitalists invite young scientists to contact them for advice.

Networking is getting easier as the number of resources connecting academia to industry continues to grow. It is no secret that the top US biotech hubs are located in academic supercentres, such as California's

Bay Area, San Diego and Boston. Increasingly, universities are providing a bridge between academia and industry (see 'Human resources'). These intellectual meccas are known as a source of fresh talent, providing a plethora of back-up opportunities.

KNOW THE MARKET

Anticipating regional market needs is important when contemplating a start-up position. Although technology-platform companies were all the rage a few years ago, most venture capitalists suggest that companies with a product are in a much better position today. "We want companies that are producing," says Stampacchia. "We avoid joining companies that just build models."

That approach is strongest in European countries, such as Germany, where venture capital has dried up. "There are fewer people willing to back companies at an early stage," says Stampacchia. Morrison predicts a painful two to three years until the market is again fruitful.

Although Scandinavia remains the bright spot in Europe, local venture capitalists suggest that the safest career move right now is to start your career in the drug industry to gain marketable experience. "For many biotech companies, you have to work according to drug-industry standards, because it is the intended customer," says Magnus Persson, a partner in HealthCap Venture Capital in Stockholm, Sweden. Indeed, many European companies are spin-offs from the drug industry. Drug development looks like a safe bet — in Europe, at least.

But Morrison doesn't advocate this approach in the United States. "Go to a young entrepreneurial life-sciences company that's doing something unique and cut your teeth there," says Morrison. "There's more opportunity in biotech."

All the experts agree that biotech is the reigning high-growth industry. And, as Alderete can attest, the risk can be worth it.

Virginia Gewin is a freelance science writer based in Portland, Oregon.

Web links

UCSF Center for BioEntrepreneurship
www.ucsf.edu/cbe
 California Institute for Quantitative Biomedical Research
www.qb3.org
 BioX at Stanford University
biox.stanford.edu
 BioSpace
www.biospace.com
 Biocom, life-science industry association
www.biocom.org
 BayBio, life-science industry association
www.baybio.com
 MassBio, life-science industry association
www.massbio.org

GRADUATE JOURNAL

Wrapping things up

After emerging from my penultimate thesis committee meeting, I experienced a range of bittersweet emotions underscored by a profound sense of nostalgia. Although I am thrilled about finishing my PhD and moving on to a postdoc at the Pasteur Institute in Paris, I will be sad to leave New York and Rockefeller University.

What I will miss most is what makes any institution special — the people. From the security guards to the full professors, there have been many who have nurtured me through graduate school. My lab mates at the Aaron Diamond AIDS Research Center and my adviser, Mark Muesing, have been like a surrogate family to me. Our relationship has been based on mutual exchange. Mark taught me the nuances of HIV biology, I taught him the finer points of rap music and hip-hop culture.

So as I wrap things up, I'll remember everyone who made my graduate experience so unforgettable. I'll take all of the knowledge and encouragement with me into my future professional experiences, and I aspire to attain a level of scientific achievement that will make them proud. In the meantime, I want them all to know that if they should crave some fine French cuisine or just some friendly conversation, I'm only a phone call or trans-Atlantic flight away. ■

Tshaka Cunningham is a fifth-year graduate student at Rockefeller University in New York.

RECRUITERS & INDUSTRY

Rules of engagement

Since the arrival of Gregory Lucier as chief executive last year, Invitrogen has been subtly broadening its role. Headquartered in Carlsbad, California, the research-tools company is seeking to build research partnerships with academic, industrial and government customers, and is focusing on the way life sciences converge with healthcare.

One significant change that Invitrogen has made is to increase its scientific recruitment. In particular, it is seeking out promising young researchers. Last year, it spent 6% of its revenue on research and development (R&D); this year it plans to spend 10% (some US\$900 million), and as a result will hire up to 250 PhD scientists over the next 12 months.

Perhaps the most aggressive tool in this recruiting drive is the company's experimental

project called the Life Science Leadership Program, which it launched in January. Under this, the firm scoured the best universities worldwide and drew up a shortlist of 15 of the brightest and best PhDs and MBA candidates. The aim was to give three of these a job.

To smooth the recruitment process, each of the candidates spent two days at the company. There they got a feel for what Invitrogen does and how it does it. More importantly, they got a chance to see the breadth of opportunities that were potentially available to them — both the range of scientific projects and the resources that back them up.

In short, the initiative tries to show how attractive Invitrogen is as a job prospect for them. They can, for example, pursue a technical career path or choose to follow a managerial route.

Ultimately, the goal is to have four 'waves' of this two-year scientific leadership programme up and running, with 20–30 people per wave. Every 6–9 months, the new recruits will be given a new assignment in a different part of the company. And after two years, they should be qualified for a fairly meaty job.

For example, Invitrogen is building an R&D centre in Frederick, Maryland, which will eventually house 150 scientists. If members of the leadership programme choose a scientific track, they could well become leaders at that facility. If they choose a management role, they could be leading teams in five years.

The key to the concept is that the programme is not a postdoc, but rather it is a 'pipeline' for middle- to senior-level scientists. ■

Joe Rodriguez, senior vice-president, human resources, Invitrogen, Carlsbad, California.

MOVERS

Eve Slater, board of directors, Vertex Pharmaceuticals, Cambridge, Massachusetts



Eve Slater's career, which includes high points in academia, industry and government, embodies the improvements that have occurred for female scientists over the past 30 years. "When you roll back the film you can see there has been progress," says Slater.

Even so, her progress was mixed with the kinds of choices that female scientists must still make today. She chose to leave the clinical practice,

research and teaching she loved in Boston — she was chief resident at Massachusetts General Hospital — because her spouse had career opportunities in New York.

Instead, she joined the drug company Merck. She began her industrial career as a basic scientist, but soon found that her ability to communicate with scientists from other specialities helped to open doors. A pivotal moment came when she made a presentation on the cholesterol drug lovastatin to the US Food and Drug Administration (FDA) for approval. After this was successful, Merck gave her a position in regulatory affairs.

Slater loved her new role, describing it as a "truly creative discipline". She enjoyed recruiting the physicians and scientists needed to answer key questions about how drugs worked in humans, and she savoured the scepticism she knew she would face at the FDA.

Her success was such that she was

considered for the post of FDA commissioner a few times. But in 2001, some influential members of Congress signed a letter saying that the FDA commissioner should never have worked for industry, which prohibited her from getting the job.

Instead, she joined the US Department of Health and Human Services as health secretary Tommy Thompson's assistant. She enjoyed the public speaking and the chance to be an advocate on many issues — especially women's health — but eventually she felt the urge to return to the private sector.

So she now fills her schedule with positions that match her interests and abilities, such as serving on boards of favourite causes and, most recently, joining Vertex as a board member. She does, however, believe that she has time and energy for "one more big job". Her creativity, adaptability and persistence are more than likely to lead her to one. ■

CV

2001–03: Assistant secretary for health, US Department of Health and Human Services

1983–2002: Adjunct associate clinical professor of medicine, Columbia University, New York

1983–2001: Merck Sharp & Dohme Research Laboratories, positions included senior vice-president of external policy, vice-president of corporate public affairs

1977–82: Chief of hypertension unit, Massachusetts General Hospital, Boston, and assistant professor of medicine, Harvard Medical School